ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 2, *pp.* 391–419

Semisupervised inference for explained variance in high dimensional linear regression and its applications

T. Tony Cai

University of Pennsylvania, Philadelphia, USA

and Zijian Guo

Rutgers University, Piscataway, USA

[Received March 2018. Final revision November 2019]

Summary. The paper considers statistical inference for the explained variance $\beta^{\mathrm{T}}\Sigma\beta$ under the high dimensional linear model $Y = X\beta + \epsilon$ in the semisupervised setting, where β is the regression vector and Σ is the design covariance matrix. A calibrated estimator, which efficiently integrates both labelled and unlabelled data, is proposed. It is shown that the estimator achieves the minimax optimal rate of convergence in the general semisupervised framework. The optimality result characterizes how the unlabelled data contribute to the estimation accuracy. Moreover, the limiting distribution for the proposed estimator is established and the unlabelled data have also proved useful in reducing the length of the confidence interval for the explained variance. The method proposed is extended to semisupervised inference for the unweighted quadratic functional $\|\beta\|_2^2$. The inference results obtained are then applied to a range of high dimensional statistical problems, including signal detection and global testing, prediction accuracy evaluation and confidence ball construction. The numerical improvement of incorporating the unlabelled data is demonstrated through simulation studies and an analysis of estimating heritability for a yeast segregant data set with multiple traits.

Keywords: Confidence set; Heritability; Minimaxity; Prediction accuracy; Signal detection; Unlabelled data

1. Introduction

High dimensional linear models are ubiquitous in contemporary statistical modelling with a wide range of applications in many scientific fields. The early focus has been mainly on developing methods for the recovery of the whole regression vector via penalized or constrained l_1-minimization approaches. Examples include the lasso (Tibshirani, 1996), Dantzig selector (Candès and Tao, 2007), minimax concave penalty (Zhang, 2010), square-root lasso (Belloni *et al.*, 2011) and scaled lasso (Sun and Zhang, 2012). There has been significant recent interest in statistical inference for low dimensional functionals, including confidence intervals and hypothesis testing for individual regression coefficients (Zhang and Zhang, 2014; van de Geer *et al.*, 2014; Javanmard and Montanari, 2014a, b), minimaxity and adaptivity of confidence intervals for general linear functionals (Cai and Guo, 2018b), estimation of the signal-to-noise-ratio (Verzelen and Gassiat, 2018; Janson *et al.*, 2017), inference for the l_q-accuracy

Address for correspondence: Zijian Guo, Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854-8019, USA.
E-mail: zijguo@stat.rutgers.edu

 1369–7412/20/82391

of a given estimator (Cai and Guo, 2018a) and estimation of quadratic functionals (Janson *et al.*, 2017; Guo *et al.*, 2019). Motivated by a range of applications, the present paper considers the semisupervised inference problem in high dimensions, where the main statistical goal is to integrate both the labelled and the unlabelled data, and to propose efficient point and interval estimators.

1.1. Problem formulation and motivation

We consider the high dimensional linear model with a random design:

$$y_i = X_{i\cdot}^{\mathrm{T}}\beta + \epsilon_i, \qquad \text{for } 1 \leqslant i \leqslant n, \tag{1}$$

where $y_i \in \mathbb{R}$ and $X_{i\cdot} \in \mathbb{R}^p$ denote respectively the outcome and the measured covariates of the ith observation, ϵ_i denotes the error and $\beta \in \mathbb{R}^p$ denotes the high dimensional regression vector. The covariates $X_{i\cdot}$ are independent and identically distributed (IID) p-dimensional random vectors with mean 0 and covariance matrix Σ and the errors $\{\epsilon_i\}_{1\leqslant i\leqslant n}$ are IID random variables with mean 0 and variance σ^2 and independent of $\{X_{i\cdot}\}_{1\leqslant i\leqslant n}$. The explained variance under the regression model (1) is represented by $Q = \mathrm{var}(X_{i\cdot}^{\mathrm{T}}\beta) = \beta^{\mathrm{T}}\Sigma\beta$. We focus on the semisupervised setting, where the data are a combination of the labelled data $\{y_i, X_{i\cdot}\}_{1\leqslant i\leqslant n}$ in the regression model (1) and the unlabelled data $\{X_{i\cdot}\}_{n+1\leqslant i\leqslant n+N}$. Here the measured covariates of both the labelled and the unlabelled data are assumed to be independent and to follow the same distribution. The more conventional supervised setting is treated as a special case with no additional unlabelled data.

The setting of semisupervised learning is commonly seen in applications where the outcomes are more expensive to collect than are the covariates. For example, in the analysis of electronic health records databases, the covariates are easy to be automatically extracted whereas labelling of the outcomes is costly and time consuming (Chakrabortty and Cai, 2018; Gronsbell and Cai, 2017). In addition, semisupervised learning naturally arises in the integrative analysis of multiple (genetics) data sets where the covariates are the same across all data sets but the outcomes that are measured vary from study to study because of the specific purposes of individual studies (van Iperen *et al.*, 2017). This can be naturally formulated as semisupervised learning, where the prespecified outcome is measured over only one or several (but not all) data sets whereas the covariates are measured across all data sets.

The construction of the optimal estimator and confidence intervals for $Q = \beta^{\mathrm{T}}\Sigma\beta$ in the semisupervised and high dimensional setting is not only of significant interest in its own right but is also closely connected to several other important statistical problems.

(a) *Heritability*: heritability is among the most important genetics concepts. Under model (1) with the outcome normalized to have unit variance, $\beta^{\mathrm{T}}\Sigma\beta$ is a measure of heritability, which quantifies the total variance explained by genetic variants (Owen, 2012; Guo *et al.*, 2019; Janson *et al.*, 2017; Verzelen and Gassiat, 2018).

(b) *Signal-to-noise ratio SNR and proportion of variance explained*: the signal-to-noise ratio SNR and proportion of variance explained are important statistical concepts and are defined respectively as $\beta^{\mathrm{T}}\Sigma\beta/(\beta^{\mathrm{T}}\Sigma\beta + \sigma^2)$ and $\beta^{\mathrm{T}}\Sigma\beta/\sigma^2$ under model (1). Together with a good estimator of σ^2 (Sun and Zhang, 2012; Belloni *et al.*, 2011), the results for $\beta^{\mathrm{T}}\Sigma\beta$ that are established in this paper are useful for inference of SNR and the proportion of variance explained.

(c) *Signal detection and global testing*: inference for the explained variance can be applied to testing the global hypothesis $H_0 : \beta = \beta^{\mathrm{null}}$ for $\beta^{\mathrm{null}} \in \mathbb{R}^p$, which includes signal detection as a special case with $\beta^{\mathrm{null}} = 0$. The connection is revealed in the adjusted linear model $y_i -$

$X_{i\cdot}^{\mathrm{T}}\beta^{\mathrm{null}} = X_{i\cdot}^{\mathrm{T}}(\beta-\beta^{\mathrm{null}}) + \epsilon_i$ for $1 \leqslant i \leqslant n$, where testing for $H_0 : \beta = \beta^{\mathrm{null}}$ is recast as testing the hypotheses $H_0 : (\beta-\beta^{\mathrm{null}})^{\mathrm{T}}\Sigma(\beta-\beta^{\mathrm{null}}) = 0$ *versus* $H_1 : (\beta-\beta^{\mathrm{null}})^{\mathrm{T}}\Sigma(\beta-\beta^{\mathrm{null}}) > 0$.

(d) *Prediction accuracy assessment*: accuracy assessment is of significant importance in applications. Let $\check{\beta}$ denote a given estimator based on the training data. We define the out-of-sample prediction accuracy for a given observation x_{new} as $\mathbb{E}_{x_{\mathrm{new}}}\{x_{\mathrm{new}}^{\mathrm{T}}(\check{\beta}-\beta)\}^2 = (\check{\beta}-\beta)^{\mathrm{T}}\Sigma(\check{\beta}-\beta)$. We introduce the following adjusted linear model for the independent test data $\{X_{i\cdot}, y_i\}_{1\leqslant i\leqslant n}$:

$$y_i - X_{i\cdot}^{\mathrm{T}}\check{\beta} = X_{i\cdot}^{\mathrm{T}}(\beta-\check{\beta}) + \epsilon_i \qquad \text{for } 1 \leqslant i \leqslant n. \tag{2}$$

Inference results developed for the explained variance can be applied to model (2) to obtain the corresponding results for the prediction accuracy $\mathbb{E}_{x_{\mathrm{new}}}\{x_{\mathrm{new}}^{\mathrm{T}}(\check{\beta}-\beta)\}^2$.

(e) *Confidence ball for* β: construction of confidence balls for β is another important application. Based on model (2), a confidence interval $(L(Z), U(Z))$ for $(\check{\beta}-\beta)^{\mathrm{T}}\Sigma(\check{\beta}-\beta)$ leads to a confidence ball for β centring at $\check{\beta}$, $\{\beta : \|\beta-\check{\beta}\|_2^2 \leqslant U(Z)/\lambda_{\min}(\Sigma)\}$, where $\lambda_{\min}(\Sigma)$ denotes the smallest eigenvalue of Σ.

More detailed discussions about these statistical applications are presented in Section 5.

1.2. Results and contributions

A central question in semisupervised learning is *how to use both labelled and unlabelled data efficiently* (Chakrabortty and Cai, 2017; Gronsbell and Cai, 2017). We introduce a novel two-step calibrated high dimensional inference for variance explained estimator called 'CHIVE', where the first step is to plug in the estimators of β and Σ, denoted by $\hat{\beta}$ and $\hat{\Sigma}$ respectively, and the second step is to calibrate this plug-in estimator $\hat{\beta}^{\mathrm{T}}\hat{\Sigma}\hat{\beta}$ through estimating a dominating term in its error decomposition. The second step is to rebalance the bias and variance and to improve the estimation accuracy. Different forms of $\hat{\beta}$ and $\hat{\Sigma}$ can be taken as inputs of the CHIVE method and this flexibility is useful in integrating the unlabelled data to estimate Σ more accurately. This idea is then extended to semisupervised inference for the unweighted quadratic functional $\|\beta\|_2^2$, where the additional unlabelled data facilitate the estimation of Σ^{-1}.

Another important question is whether the unlabelled data have been efficiently utilized in semisupervised learning. We address this question by establishing the minimax optimal rate of convergence for estimating $\beta^{\mathrm{T}}\Sigma\beta$, where the optimal rate is $M/\sqrt{n} + M^2/\sqrt{(N+n)} + k\log(p)/n$, with p, n, N, k and M denoting respectively the dimension, the size of the labelled data, the size of the unlabelled data, the sparsity and the l_2-norm of β. The proposed CHIVE estimator achieves this optimal rate, which justifies the efficient use of the unlabelled data. The optimal rate is not just achieved for the case where there is a large amount of unlabelled data but is also for any given amount of unlabelled data. The minimax optimal rate characterizes the fundamental difficulty of the inference problem in the semisupervised setting and is independent of specific procedures. This minimax rate also reveals that the unlabelled data are most effective when the signal strength $\|\beta\|_2$ is large.

We establish the limiting distribution of the CHIVE estimator and construct data-driven confidence intervals for $\beta^{\mathrm{T}}\Sigma\beta$ based on this estimator. The limiting distribution is normal and its variance is scaled to the proportion of the labelled data, which is unique to the semisupervised setting. A larger amount of unlabelled data leads to a smaller proportion of the labelled data and hence a smaller asymptotic variance, which leads to a shorter confidence interval for $\beta^{\mathrm{T}}\Sigma\beta$. The effect of the unlabelled data is also demonstrated in the numerical studies. Specifically, in comparison with the estimators based only on the labelled data, the root-mean-squared

estimator RMSE for estimation and the length of confidence intervals can be reduced by as much as 70%. See the details in Section 6.

The improvement in semisupervised inference for $\|\beta\|_2^2$ is similar to that for $\beta^T\Sigma\beta$ at a high level but different in technical details. Specifically, the estimation accuracy is significantly improved in the strong signal regime, and the improvement is limited if the signal strength $\|\beta\|_2^2$ is not sufficiently large. Construction of confidence intervals for $\|\beta\|_2^2$ also becomes easier in the sense that the condition for sample size and model complexity is weakened by making use of the unlabelled data.

The inference results that are obtained in this paper are applied to

(a) signal detection and global testing,
(b) prediction accuracy evaluation and
(c) confidence ball construction.

For signal detection, we control the type I error and characterize the type II error by establishing the power function under a local alternative. The results can be easily extended to the general global testing problem. For evaluation of out-of-sample prediction accuracy of a given sparse estimator of β, both the point and the interval estimators are developed. We establish the estimation error bound for the point estimator of the prediction accuracy and control the length of the corresponding confidence interval. A confidence ball for the regression vector β with controlled radius is also constructed. We stress that these procedures are data driven and do not require *a priori* knowledge of the design covariance matrix Σ or the noise level σ. See more details in Section 5.

1.3. Related work

Estimation and inference for quadratic functionals have been studied in the literature in a range of settings. In particular, minimax and adaptive estimation of quadratic functionals plays an important role in non-parametric inference and has been well studied in density estimation, non-parametric regression and the white noise with drift model. See, for example, Bickel and Ritov (1988), Donoho and Nussbaum (1990), Efromovich and Low (1996), Laurent and Massart (2000), Cai and Low (2005, 2006) and Collier *et al.* (2017).

The most related works to the current paper are Verzelen and Gassiat (2018) and Guo *et al.* (2019), which considered estimation of $\beta^T\Sigma\beta/\sigma^2$ and $\|\beta\|_2^2$ respectively, in high dimensional linear regression. The main difference between the current paper and these two related works are twofold.

(a) Verzelen and Gassiat (2018) and Guo *et al.* (2019) considered only the supervised setting instead of the semisupervised setting. As demonstrated in both theoretical and numerical justifications, a careful integration of the unlabelled data proves useful in improving the estimation accuracy and reducing the length of constructed confidence intervals.
(b) The focus of Verzelen and Gassiat (2018) and Guo *et al.* (2019) is about point estimation whereas the current paper studies the more challenging problem of uncertainty quantification and also related hypothesis testing, in addition to point estimation. As is well known, uncertainty quantification in high dimensions is significantly different from and more involved than point estimation (Nickl and van de Geer, 2013; Cai and Guo, 2017).

Another related reference, Janson *et al.* (2017), studied the construction of confidence intervals for $\|\beta\|_2^2$ in the setting of $\Sigma = I$, moderate dimension, where $n/p \to \xi \in (0, 1)$ and no sparsity assumption on β. The inference problem that is considered in the current paper is significantly different from the setting that was considered in Janson *et al.* (2017), mainly because of the

complicated geometry that is induced by the sparsity structure and the unknown design covariance matrix Σ. Other works that are related to quadratic functional inference include the construction of confidence intervals for the l_2-loss of the estimator that was considered in Cai and Guo (2018b). In addition, Javanmard and Lee (2017) and Zhu and Bradic (2017) considered hypothesis testing for high dimensional linear regression. As another significant difference, the current paper studies how to integrate the labelled and unlabelled data efficiently in the general semisupervised setting whereas all the aforementioned works solely focused on supervised regression.

The statistical applications that are studied in this paper have also been considered separately in the literature. Signal detection was studied in Ingster *et al.* (2010) and Arias-Castro *et al.* (2011) under the linear model (1) in a special setting where the design covariance matrix Σ is equal to or close to the identity matrix. In this setting, Ingster *et al.* (2010) and Arias-Castro *et al.* (2011) established an optimal signal detection method and theory. The results that are established in the present paper enable the study of signal detection under a general setting where the design covariance matrix Σ is unknown. The confidence ball construction for the whole regression vector was considered in Nickl and van de Geer (2013) in the case of known σ and the optimal size and possibility of adaptive confidence balls was also established. The results that are obtained in the current paper lead to a confidence ball construction for β in the case of unknown σ. A problem that is related to prediction accuracy is inference for the estimation accuracy, which was considered in Cai and Guo (2018b) and Janson *et al.* (2017). However, inference for the prediction accuracy and that for the estimation accuracy are different problems.

1.4. Organization of the paper

The rest of the paper is organized as follows. In Section 2, we introduce in detail the CHIVE estimator and establish its minimax rate optimality in the semisupervised setting. Section 3 focuses on quantifying the uncertainty of the CHIVE estimator and construction of the confidence intervals for $\beta^{\mathrm{T}}\Sigma\beta$. In Section 4, we extend the methodology to semisupervised inference for $\|\beta\|_2^2$. We apply in Section 5 the procedures developed to tackle three important problems: signal detection and global testing, prediction accuracy evaluation and confidence ball construction. Simulation results are given in Section 6 to illustrate the numerical improvement through incorporating the unlabelled data. An analysis of a yeast data set is presented in Section 7. A discussion is provided in Section 8. The proofs and the additional simulation results are presented in the on-line appendix.

2. Semisupervised estimation of $\beta^{\mathrm{T}}\Sigma\beta$

In this section, we first introduce the calibration methodology for estimating the variance explained in the general semisupervised framework and then establish the minimax convergence rate of estimating $\beta^{\mathrm{T}}\Sigma\beta$. A significant statistical gain is obtained by carefully integrating the unlabelled data and the estimator proposed is shown to achieve the optimal rate in the semisupervised setting. The supervised setting and the setting with known design covariance matrix are then discussed as special cases. We begin with the notation that will be used in the rest of the paper.

For a matrix A, $A_{i\cdot}$, $A_{\cdot j}$ and $A_{i,j}$ denote respectively the ith row, jth column and (i, j) entry of the matrix A. The spectral norm of A is $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$ and the matrix l_1-norm is $\|A\|_{L_1} = \sup_{1\leqslant j\leqslant p} \Sigma_{i=1}^{p} |A_{ij}|$. For a symmetric matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote respectively the smallest and largest eigenvalue of A. For a set S, $|S|$ denotes the cardinality of S. For a vector $x \in \mathbb{R}^p$, $\mathrm{supp}(x)$ denotes the support of x and the l_q-norm of x is defined

as $\|x\|_q = (\Sigma_{i=1}^p |x_i|^q)^{1/q}$ for $q \geqslant 0$ with $\|x\|_0 = |\mathrm{supp}(x)|$ and $\|x\|_\infty = \max_{1 \leqslant j \leqslant p} |x_j|$. For $a \in \mathbb{R}$, $a_+ = \max\{a, 0\}$. We use c and C to denote generic positive constants that may vary from place to place. For a sequence of random variables X_n indexed by n, we use $X_n \to^{\mathrm{p}} X$ and $X_n \to^{\mathrm{d}} X$ to represent that X_n converges to X in probability and in distribution respectively. For a sequence of random variables X_n and numbers a_n, we define $X_n = o_p(a_n)$ if X_n/a_n converges to 0 in probability. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means that $a_n \leqslant Cb_n$ for all n and $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\overline{\lim}_{n\to\infty} a_n/b_n = 0$ and $a_n \gg b_n$ if $b_n \ll a_n$. We define the signal-to-noise ratio SNR in the context of model (1) as $\mathrm{SNR} = (1/\sigma)\sqrt{(\beta^{\mathrm{T}}\Sigma\beta)}$.

2.1. Calibration of plug-in estimators

In semisupervised learning, we observe the labelled data $(X_{1\cdot}, y_1), \ldots, (X_{n\cdot}, y_n)$ and the unlabelled data $X_{n+1\cdot}, \ldots, X_{n+N\cdot}$, where $X_{1\cdot}, \ldots, X_{n\cdot}, X_{n+1\cdot}, \ldots, X_{n+N\cdot}$ are IID realizations of p-dimensional covariates. We use $\hat{\beta}$ and $\hat{\Sigma}$ to denote some estimators of β and Σ, which will be specified later. A preliminary estimator of the quadratic functional $Q = \beta^{\mathrm{T}}\Sigma\beta$ is the plug-in estimator $\hat{\beta}^{\mathrm{T}}\hat{\Sigma}\hat{\beta}$, which has the error decomposition

$$\hat{\beta}^{\mathrm{T}}\hat{\Sigma}\hat{\beta} - \beta^{\mathrm{T}}\Sigma\beta = 2\hat{\beta}^{\mathrm{T}}\hat{\Sigma}(\hat{\beta} - \beta) - (\hat{\beta} - \beta)^{\mathrm{T}}\hat{\Sigma}(\hat{\beta} - \beta) + \beta^{\mathrm{T}}(\hat{\Sigma} - \Sigma)\beta. \tag{3}$$

Since the first term $2\hat{\beta}^{\mathrm{T}}\hat{\Sigma}(\hat{\beta} - \beta)$ on the right-hand side can be estimated in a data-dependent way, the corresponding estimation error of the preliminary estimator $\hat{\beta}^{\mathrm{T}}\hat{\Sigma}\hat{\beta}$ can be further reduced. We estimate the term $2\hat{\beta}^{\mathrm{T}}\hat{\Sigma}(\hat{\beta} - \beta)$ by $-2\hat{\beta}^{\mathrm{T}}(1/n)\Sigma_{i=1}^n X_{i\cdot}(y_i - X_{i\cdot}\hat{\beta})$ and propose the following calibrated estimator:

$$\hat{Q}(\hat{\beta}, \hat{\Sigma}) = \hat{\beta}^{\mathrm{T}}\hat{\Sigma}\hat{\beta} + 2\hat{\beta}^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^n X_{i\cdot}(y_i - X_{i\cdot}\hat{\beta}). \tag{4}$$

This estimator is referred to as the calibrated high dimensional inference for variance explained estimator CHIVE. The calibration step in equation (4) is essentially to improve the plug-in estimator $\hat{\beta}^{\mathrm{T}}\hat{\Sigma}\hat{\beta}$ through rebalancing the bias and variance.

The CHIVE estimator requires three inputs: the initial estimators $\hat{\beta}$ and $\hat{\Sigma}$ and the data (X, y). With this machinery, we have the flexibility of choosing the initial estimators $\hat{\beta}$ (and also $\hat{\sigma}^2$) and $\hat{\Sigma}$ on the basis of the observed data. We begin with the estimator for β and σ^2 and then move on to the estimator for Σ. Throughout the paper, we assume that the estimators $\hat{\beta}$ and $\hat{\sigma}^2$ satisfy the following conditions.

Condition 1. With probability larger than $1 - \gamma(n)$ where $\gamma(n) \to 0$, the estimator $\hat{\beta}$ satisfies

$$\max\left\{\frac{1}{n}\sum_{i=1}^n \{X_{i\cdot}^{\mathrm{T}}(\hat{\beta} - \beta)\}^2, \|\hat{\beta} - \beta\|_2^2\right\} \lesssim \frac{k\log(p)}{n}\sigma, \qquad \|(\hat{\beta} - \beta)_{S^{\mathrm{c}}}\|_1 \leqslant C_0\|(\hat{\beta} - \beta)_S\|_1$$

where $S = \mathrm{supp}(\beta)$ and $C_0 > 0$ is some positive constant.

Condition 2. $\hat{\sigma}^2$ is a consistent estimator of σ^2, i.e. $|\hat{\sigma}^2/\sigma^2 - 1| \to^{\mathrm{p}} 0$.

One of the key assumptions for the general penalized estimators satisfying conditions 1 and 2 is the following restricted eigenvalue condition on the population covariance matrix Σ:

$$\kappa(k, C_0, \Sigma) = \min_{S \in \{1,\ldots,p\}, |S| \leqslant k} \min_{v \neq 0, \|v_{S^{\mathrm{c}}}\|_1 \leqslant C_0\|v_S\|_1} \frac{\|\Sigma^{1/2}v\|_2}{\|v_S\|_2} \geqslant c,$$

for some positive constant $c > 0$. This population version restricted eigenvalue condition implies the sample version restricted eigenvalue condition that was introduced in Bickel *et al.* (2009),

under the assumption that the covariates $X_{i\cdot}$ are in a certain broad family of sub-Gaussian random vectors and the sparsity k satisfies $k \lesssim n/\log(p)$; see Zhou (2009) and Raskutti *et al.* (2010) for the exact statement.

2.1.1. Estimators satisfying conditions 1 and 2

The scaled lasso estimator $\{\hat{\beta}, \hat{\sigma}\}$ that is defined by

$$\{\hat{\beta}, \hat{\sigma}\} = \arg \min_{\beta \in \mathbb{R}^p, \sigma \in \mathbb{R}^+} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \sqrt{\left\{ \frac{2.01 \log(p)}{n} \right\}} \sum_{j=1}^{p} \frac{\|X_{\cdot j}\|_2}{\sqrt{n}} |\beta_j| \tag{5}$$

has been shown in Sun and Zhang (2012) to satisfy conditions 1 and 2 under regularity conditions. See also lemma 1 in Guo *et al.* (2019) for more details. Since the square-root lasso estimator (Belloni *et al.*, 2011) is numerically the same as the scaled lasso estimator, the square-root lasso estimators of β and σ also satisfy conditions 1 and 2. In addition, with *a priori* knowledge of σ, the lasso estimator of β and other variants are also shown to satisfy condition 1; see Candès and Tao (2007), Zhang (2010) and Ye and Zhang (2010) for more details.

Now, we turn to the estimators of Σ. This is exactly the place where we make use of the unlabelled data. Specifically, we pool the information that is contained in both the labelled and the unlabelled data and estimate Σ by

$$\hat{\Sigma}^{\mathrm{S}} = \frac{1}{n+N} \sum_{i=1}^{n+N} X_{i\cdot} X_{i\cdot}^{\mathrm{T}}.$$

Then we use $\hat{\beta}$ and $\hat{\Sigma}^{\mathrm{S}}$ as inputs and utilize the calibration idea that was introduced in equation (4):

$$\hat{Q}(\hat{\beta}, \hat{\Sigma}^{\mathrm{S}}) = \hat{\beta}^{\mathrm{T}} \hat{\Sigma}^{\mathrm{S}} \hat{\beta} + 2\hat{\beta}^{\mathrm{T}} \frac{1}{n} \sum_{i=1}^{n} X_{i\cdot}(y_i - X_{i\cdot}^{\mathrm{T}} \hat{\beta}). \tag{6}$$

When there is no confusion, we use $\hat{Q}$ to denote the estimator that is proposed in equation (6). We introduce the following regularity conditions and then establish the convergence rate of the proposed estimator in equation (6) in theorem 1.

Assumption 1. The regression vector β is assumed to be k sparse; the errors $\{\epsilon_i\}_{1 \leqslant i \leqslant n}$ are independent of $\{X_{i\cdot}\}_{1 \leqslant i \leqslant n+N}$ and follow IID sub-Gaussian random variables with mean 0 and variance σ^2; the rows $X_{i\cdot}$ are IID p-dimensional random vectors and can be expressed in the form of $X_{i\cdot} = \Sigma^{1/2} Z_{i\cdot}$ where $Z_{i\cdot} \in \mathbf{R}^p$ is a sub-Gaussian random vector of mean 0 and identity covariance matrix and Σ has a bounded restricted largest eigenvalue $\rho_{\max}(k, \Sigma)$, which is defined as $\rho_{\max}(k, \Sigma) = \max_{\|v\|_2 = 1, \|v\|_0 \leqslant k} v^{\mathrm{T}} \Sigma v$.

Assumption 2. $\sqrt{\mathbb{E}(\beta^{\mathrm{T}} X_{1\cdot} X_{1\cdot}^{\mathrm{T}} \beta - \beta^{\mathrm{T}} \Sigma \beta)^2} \geqslant c_0 \beta^{\mathrm{T}} \Sigma \beta$, for some positive constant $c_0 > 0$.

Assumption 1 requires that the restricted largest eigenvalue $\rho_{\max}(k, \Sigma)$ is upper bounded, where 'restricted' here means that the maximum in the definition of $\rho_{\max}(k, \Sigma)$ is taken with respect to k-sparse vectors. Note that the restricted (smallest) eigenvalue condition is not required for the theoretical analysis of the proposed estimator $\hat{Q}$ as long as the estimator $\hat{\beta}$ of β satisfies condition 1. Define $U = X_{i\cdot}^{\mathrm{T}} \beta / \sqrt{(\beta^{\mathrm{T}} \Sigma \beta)}$, where $\mathbb{E}(U) = 0$ and $\mathbb{E}(U^2) = 1$. Assumption 2 is placed on this random variable U such that $\mathrm{var}(U^2)$ is not vanishing. This assumption is imposed such that $\mathrm{var}(U^2)$ can be well estimated and this type of assumption has been introduced in the covariance matrix estimation literature (Cai and Liu, 2011) for the same purpose.

Theorem 1. Suppose that assumption 1 holds and $k \leqslant cn/\log(p)$ for some constant $c > 0$. For any estimator $\hat{\beta}$ satisfying condition 1, with probability at least $1 - \gamma(n) - C\{p^{-c} +$

$\exp(-cN)+\exp(-ct^2)\}$, the estimator $\hat{Q}=\hat{Q}(\hat{\beta},\hat{\Sigma}^{\mathrm{S}})$ defined in equation (6) satisfies

$$|\hat{Q}-Q|\lesssim t\frac{\sigma\|\Sigma^{1/2}\beta\|_2}{\sqrt{n}}+t\frac{\beta^{\mathrm{T}}\Sigma\beta}{\sqrt{(N+n)}}+\left(1+\frac{\|\Sigma^{1/2}\beta\|_2}{\sigma}\frac{N}{n+N}\right)\frac{k\log(p)}{n}\sigma^2. \tag{7}$$

Under the additional assumptions $k\ll\sqrt{n}/\log(p)$ and $\mathrm{SNR}\gg k\log(p)/\sqrt{n}$,

$$\frac{\sqrt{n}(\hat{Q}-Q)}{\sqrt{\{4\sigma^2\beta^{\mathrm{T}}\Sigma\beta+\rho\mathbb{E}(\beta^{\mathrm{T}}X_1.X_{1.}^{\mathrm{T}}\beta-\beta^{\mathrm{T}}\Sigma\beta)^2\}}}\xrightarrow{\mathrm{d}}N(0,1) \tag{8}$$

where $\rho=\lim_{n\to\infty}n/(N+n)$.

As a remark, the probability $1-\gamma(n)-C\{p^{-c}+\exp(-cN)+\exp(-ct^2)\}$ holds for the finite sample n and finite dimension p and also any non-negative constant $t\geqslant 0$. However, the established result is more interesting over the regime $\min\{p,n\}\to\infty$ and $t\to\infty$ as, in this scenario, the corresponding probability $1-\gamma(n)-C\{p^{-c}+\exp(-cN)+\exp(-ct^2)\}$ approaches 1. Since $Q\geqslant 0$, the convergence rate (7) also holds for $\hat{Q}_+$: the positive part of $\hat{Q}$. To keep the notation simpler, we present only the results for $\hat{Q}$ in this paper.

The rate of convergence in expression (7) reveals the effect of the unlabelled data. The sample size of the unlabelled data, N, appears only in the second term $t\beta^{\mathrm{T}}\Sigma\beta/\sqrt{(N+n)}$. An interesting observation is that the usefulness of the unlabelled data varies across different signal strengths. If the signal is strong in the sense that $\mathrm{SNR}\gtrsim\max\{1,k\log(p)/\sqrt{n}\}$, in which case the term $t\beta^{\mathrm{T}}\Sigma\beta/\sqrt{(N+n)}$ is dominant in expression (7), then the additional unlabelled data reduce the rate of convergence significantly; if the signal is weak in the sense that $\mathrm{SNR}\ll\max\{1,k\log(p)/\sqrt{n}\}$, then the effect of the additional unlabelled data is limited.

To demonstrate the effect of calibration, we note that an upper bound for the term $\hat{\beta}^{\mathrm{T}}\hat{\Sigma}(\hat{\beta}-\beta)$ in expression (3) is of the order of magnitude $\sigma\|\Sigma^{1/2}\beta\|_2\sqrt{\{k\log(p)/n\}}$ whereas the remaining error after the calibration step is

$$t\frac{\sigma\|\Sigma^{1/2}\beta\|_2}{\sqrt{n}}+\left(1+\frac{\|\Sigma^{1/2}\beta\|_2}{\sigma}\frac{N}{n+N}\right)\frac{k\log(p)}{n}\sigma^2,$$

as shown in expression (7). By comparing these upper bounds, we note that the calibration step is useful in reducing the upper bound for the rate of convergence. This reduction of estimation error is also numerically demonstrated in Section 6.2. The terms

$$t\frac{\beta^{\mathrm{T}}\Sigma\beta}{\sqrt{(N+n)}}+\frac{k\log(p)}{n}\sigma^2$$

in expression (7) capture the convergence rate of the last two terms in expression (3).

The distributional result in expression (8) is established under the additional assumptions $k\ll\sqrt{n}/\log(p)$ and $\mathrm{SNR}\gg k\log(p)/\sqrt{n}$. These additional assumptions are imposed to ensure that the variance component

$$t\frac{\sigma\|\Sigma^{1/2}\beta\|_2}{\sqrt{n}}+t\frac{\beta^{\mathrm{T}}\Sigma\beta}{\sqrt{(N+n)}},$$

captured by the normal limiting distribution after rescaling, dominates the bias component

$$\left(1+\frac{\|\Sigma^{1/2}\beta\|_2}{\sigma}\frac{N}{n+N}\right)\frac{k\log(p)}{n}\sigma^2.$$

Since the bias term is difficult to characterize, we impose these sufficient conditions such that the variance term is the dominating term. The normal limiting distribution in expression (8) can be used in Section 3 to construct confidence intervals for $\beta^{\mathrm{T}}\Sigma\beta$.

Another interesting phenomenon is that the limiting distribution that is established in expression (8) depends on the proportion of the labelled data, which is unique in the semisupervised inference problem. If the amount of unlabelled data dominates that of labelled data (i.e. $\rho = 0$), then the limiting distribution in expression (8) is simplified to

$$\frac{\sqrt{n}(\hat{Q} - Q)}{\sqrt{(4\sigma^2\beta^{\mathrm{T}}\Sigma\beta)}} \xrightarrow{\mathrm{d}} N(0, 1).$$

Theorem 1 demonstrates that the CHIVE estimator integrating the unlabelled data improves the rate of convergence in estimating the explained variance. The lower bound that is given in the next subsection shows that CHIVE is optimal in terms of the rate of convergence.

2.2. Optimal estimation in the semisupervised setting

In this section, we further investigate the fundamental limit for estimating $Q = \beta^{\mathrm{T}}\Sigma\beta$ in the general semisupervised setting over the specific parameter space

$$\Theta(k, M) = \{\theta = (\beta, \Sigma, \sigma) : \|\beta\|_0 \leqslant k,\ M/2 \leqslant \|\beta\|_2 \leqslant M,\ 1/M_1 \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant M_1,\ \sigma \leqslant M_2\},$$

where $M_1 \geqslant 1$ and $M_2 > 0$ are positive constants. Here k quantifies the sparsity of β and M quantifies the signal strength of the true signal β in terms of its l_2-norm. Both k and M are allowed to grow with n and p. The other conditions $1/M_1 \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant M_1$ and $\sigma \leqslant M_2$ are regularity conditions. The following theorem establishes the minimax lower bounds for estimating Q over the parameter space $\Theta(k, M)$.

Theorem 2. Suppose that $k \leqslant c\min\{n/\log(p), p^{\nu}\}$ for some constants $c > 0$ and $0 \leqslant \nu < \frac{1}{2}$. Then

$$\inf_{\tilde{Q}} \sup_{\theta \in \Theta(k,M)} \mathbb{P}\left[|\tilde{Q} - Q| \gtrsim \frac{M^2}{\sqrt{(N+n)}} + \min\left\{\frac{M}{\sqrt{n}} + \frac{k\log(p)}{n}, M^2\right\}\right] \geqslant \frac{1}{4}. \tag{9}$$

One interesting observation of theorem 2 is that only the first term in the lower bound is involved with the amount of the unlabelled data. Theorems 1 and 2 together show that the estimator that was proposed in Section 2.1 is minimax rate optimal under regularity conditions.

Corollary 1. Suppose that assumption 1 holds and $k \leqslant c\min\{n/\log(p), p^{\nu}\}$ for some constants $c > 0$ and $0 \leqslant \nu < \frac{1}{2}$. For any estimator $\hat{\beta}$ satisfying condition 1, the estimator $\hat{Q}$ defined in equation (6) is minimax rate optimal over $\Theta(k, M)$ where $\sqrt{\{k\log(p)/n\}} \lesssim M \leqslant C$ for some constant $C > 0$, i.e.

$$\sup_{\theta \in \Theta(k,M)} \mathbb{P}\left\{|\hat{Q} - Q| \gtrsim t\frac{M^2}{\sqrt{(n+N)}} + \frac{M}{\sqrt{n}} + \frac{k\log(p)}{n}\right\} \leqslant C\{p^{-c} + \exp(-cN) + \exp(-ct^2)\} + \gamma(n) \tag{10}$$

The CHIVE estimator attains the optimal convergence rate when the l_2-norm of β is relatively strong, i.e. M is bounded away from zero by $\sqrt{\{k\log(p)/n\}}$. As shown in theorem 2, for the case where $M \ll \sqrt{\{k\log(p)/n\}}$, the lower bound of estimating $\beta^{\mathrm{T}}\Sigma\beta$ is M^2. This optimal convergence rate can be achieved by a trivial estimator 0.

In corollary 1, the lower bound (9) is only matched for the regime where $M \leqslant C$ for some constant $C > 0$. For theoretical interest, we shall modify the proposed estimator $\hat{Q}$ defined in equation (6) such that the modified version achieves the lower bound (9) over the regime $M \gtrsim \sqrt{\{k \log(p)/n\}}$. We randomly split the data (y, X) into two subsamples $(y^{(1)}, X^{(1)})$ with sample size n_1 and $(y^{(2)}, X^{(2)})$ with sample size n_2, where $n_1 \asymp n_2$. Let $\hat{\beta}$ denote an estimator which is produced by the first subsample $(y^{(1)}, X^{(1)})$ and satisfies assumption 1. One example of such an estimator is the scaled lasso estimator (5) applied to the subsample $(y^{(1)}, X^{(1)})$. We propose the following estimator of Q:

$$\hat{Q}(\hat{\beta}, \hat{\Sigma}^{(2)}) = \hat{\beta}^{\mathrm{T}} \hat{\Sigma}^{(2)} \hat{\beta} + 2\hat{\beta}^{\mathrm{T}} \frac{1}{n_2} \sum_{i=n_1+1}^{n} X_{i\cdot}^{\mathrm{T}} (y_i - X_{i\cdot} \hat{\beta}), \tag{11}$$

where

$$\hat{\Sigma}^{(2)} = \frac{1}{n + N - n_1} \sum_{i=n_1+1}^{n+N} X_{i\cdot} X_{i\cdot}^{\mathrm{T}}.$$

The following theorem establishes the convergence rate of $\hat{Q}(\hat{\beta}, \hat{\Sigma}^{(2)})$ and shows that this estimator achieves the optimal convergence rate of estimating Q for $M \gtrsim \sqrt{\{k \log(p)/n\}}$.

Theorem 3. Suppose that condition 1 holds and $k \leqslant cn/\log(p)$ for some constant $c > 0$. Let $\hat{\beta}$ be an estimator depending on the first half-sample $(y^{(1)}, X^{(1)})$ and satisfying assumption 1. Then, with probability larger than $1 - \gamma(n) - C\{p^{-c} + \exp(-cN) + \exp(-ct^2)\}$, the estimator $\hat{Q}(\hat{\beta}, \hat{\Sigma}^{(2)})$ defined in equation (11) satisfies

$$|\hat{Q}(\hat{\beta}, \hat{\Sigma}^{(2)}) - Q| \lesssim (t+1) \frac{\sigma \|\Sigma^{1/2} \beta\|_2}{\sqrt{n}} + t \frac{\beta^{\mathrm{T}} \Sigma \beta}{\sqrt{(N+n)}} + \frac{k \log(p)}{n} \sigma^2. \tag{12}$$

Hence, the estimator $\hat{Q}(\hat{\beta}, \hat{\Sigma}^{(2)})$ defined in equation (11) achieves the optimal estimation rate over $\Theta(k, M)$ in the sense of inequality (10) over the regime $k \leqslant c \min\{n/\log(p), p^{\nu}\}$ for some constants $c > 0$ and $0 \leqslant \nu < \frac{1}{2}$ and $M \gtrsim \sqrt{\{k \log(p)/n\}}$.

2.3. Two special cases

We now turn to two important special cases: the inference in the supervised setting and the setting with known design covariance matrix.

2.3.1. Case I: supervised inference

In the supervised setting without any additional unlabelled data, Σ is estimated by $\hat{\Sigma}^{\mathrm{L}} = (1/n) \Sigma_{i=1}^{n} X_{i\cdot} X_{i\cdot}^{\mathrm{T}}$. The following corollary establishes the rate of convergence of the estimator $\hat{Q} = \hat{Q}(\hat{\beta}, \hat{\Sigma}^{\mathrm{L}})$, which is a special case of the estimator (6) with $N = 0$.

Corollary 2. Suppose that assumption 1 holds and $k \leqslant cn/\log(p)$ for some constant $c > 0$. For any estimator $\hat{\beta}$ satisfying condition 1, with probability larger than $1 - \gamma(n) - C\{p^{-c} + \exp(-ct^2)\}$, $\hat{Q}(\hat{\beta}, \hat{\Sigma}^{\mathrm{L}})$ proposed in expression (4) with $\hat{\Sigma}^{\mathrm{L}} = (1/n) \Sigma_{i=1}^{n} X_{i\cdot} X_{i\cdot}^{\mathrm{T}}$ satisfies

$$|\hat{Q}(\hat{\beta}, \hat{\Sigma}^{\mathrm{L}}) - Q| \lesssim t \frac{\sigma \|\Sigma^{1/2} \beta\|_2 + \beta^{\mathrm{T}} \Sigma \beta}{\sqrt{n}} + \frac{k \log(p)}{n} \sigma^2. \tag{13}$$

Under the additional assumption 2 and $\mathrm{SNR} \gg \min[k \log(p)/\sqrt{n}, \{k \log(p)/\sqrt{n}\}^{1/2}]$,

$$\frac{\sqrt{n}\{\hat{Q}(\hat{\beta},\hat{\Sigma}^{\mathrm{L}})-Q\}}{\sqrt{\{4\sigma^2\beta^{\mathrm{T}}\Sigma\beta+\mathbb{E}(\beta^{\mathrm{T}}X_{1\cdot}X_{1\cdot}^{\mathrm{T}}\beta-\beta^{\mathrm{T}}\Sigma\beta)^2\}}}\xrightarrow{\mathrm{d}}N(0,1). \tag{14}$$

Corollary 2 basically follows from theorem 1 with $N=0$ except for some technical difference. By comparing corollary 2 with theorems 1 and 3, we observe that the unlabelled data lead to a faster convergence rate by reducing $\beta^{\mathrm{T}}\Sigma\beta/\sqrt{n}$ in expression (13) to $\beta^{\mathrm{T}}\Sigma\beta/\sqrt{(N+n)}$ in expressions (7) and (12); the unlabelled data do not affect other terms in the rate of convergence. The effect of the unlabelled data is also revealed in the limiting distribution in expression (14), where the exact variance level is reduced from $\{4\sigma^2\beta^{\mathrm{T}}\Sigma\beta+\mathbb{E}(\beta^{\mathrm{T}}X_{1\cdot}X_{1\cdot}^{\mathrm{T}}\beta-\beta^{\mathrm{T}}\Sigma\beta)^2\}/n$ in expression (14) to $\{4\sigma^2\beta^{\mathrm{T}}\Sigma\beta+\rho\mathbb{E}(\beta^{\mathrm{T}}X_{1\cdot}X_{1\cdot}^{\mathrm{T}}\beta-\beta^{\mathrm{T}}\Sigma\beta)^2\}/n$ in expression (8) for $\rho=\lim_{n\to\infty}n/(N+n)\in[0,1]$. The following corollary further establishes the minimax rate for estimating $\beta^{\mathrm{T}}\Sigma\beta$ in the supervised setting.

Corollary 3. Suppose that assumption 1 holds and $k\leqslant c\min\{n/\log(p),p^{\nu}\}$ for some constants $c>0$ and $0\leqslant\nu<\frac{1}{2}$. For any estimator $\hat{\beta}$ satisfying condition 1, the estimator $\hat{Q}=\hat{Q}(\hat{\beta},\hat{\Sigma}^{\mathrm{L}})$ defined in expression (4) with $\hat{\Sigma}^{\mathrm{L}}=(1/n)\Sigma_{i=1}^{n}X_{i\cdot}X_{i\cdot}^{\mathrm{T}}$ achieves the optimal estimation rate over $\Theta(k,M)$ for $M\gtrsim\sqrt{\{k\log(p)/n\}}$, i.e. $\hat{Q}(\hat{\beta},\hat{\Sigma}^{\mathrm{L}})$ satisfies

$$\sup_{\theta\in\Theta(k,M)}\mathbb{P}\left\{|\hat{Q}(\hat{\beta},\hat{\Sigma}^{\mathrm{L}})-Q|\gtrsim t\frac{M^2}{\sqrt{n}}+\frac{M}{\sqrt{n}}+\frac{k\log(p)}{n}\right\}\leqslant C\{p^{-c}+\exp(-ct^2)\}+\gamma(n). \tag{15}$$

Remark 1. In the supervised setting, Guo *et al.* (2019) established that the optimal rate of estimating $\|\beta\|_2^2$ over $\Theta(k,M)$ for $M\gtrsim\sqrt{\{k\log(p)/n\}}$ is $M/\sqrt{n}+(M+1)k\log(p)/n$. In contrast with inequality (15), we can see that neither of these two problems is easier than the other, where there is an additional term $M^2/\sqrt{n}$ in inequality (15) and an additional term $Mk\log(p)/n$ in the optimal convergence rate of estimating $\|\beta\|_2^2$.

Inference for $\beta^{\mathrm{T}}\Sigma\beta$ in the supervised setting is closely connected to Sun and Zhang (2012) and Verzelen and Gassiat (2018), where Sun and Zhang (2012) studied the inference problem for σ^2 and Verzelen and Gassiat (2018) studied the estimation of $\beta^{\mathrm{T}}\Sigma\beta/\sigma^2$. In particular, Sun and Zhang (2012) proposed the scaled lasso estimator $\hat{\sigma}^2$ in expression (5) to estimate σ^2 and Verzelen and Gassiat (2018) proposed to estimate $\beta^{\mathrm{T}}\Sigma\beta$ by $((1/n)\|y\|_2^2-\hat{\sigma}^2)_+$ as an intermediate step of estimating $\beta^{\mathrm{T}}\Sigma\beta/\sigma^2$. For the estimator $\hat{Q}(\hat{\beta},\hat{\Sigma}^{\mathrm{L}})$ defined in equation (4), if $\hat{\beta}$ is taken as the scaled lasso estimator, then $\hat{Q}(\hat{\beta},\hat{\Sigma}^{\mathrm{L}})$ is reduced to being the same as the estimator that was proposed in Verzelen and Gassiat (2018), where the equivalence is shown by the expression

$$\hat{\beta}^{\mathrm{T}}\hat{\Sigma}^{\mathrm{L}}\hat{\beta}+2\hat{\beta}^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}(y_i-X_{i\cdot}\hat{\beta})=\frac{1}{n}(\|y\|_2^2-\|y-X\hat{\beta}\|_2^2)=\frac{1}{n}\|y\|_2^2-\hat{\sigma}^2. \tag{16}$$

As a remark, in the supervised setting, the calibration idea in expression (4) provides a completely new perspective on estimation of $\beta^{\mathrm{T}}\Sigma\beta$, where, instead of using the expression $Q=\mathbb{E}(y_i^2)-\sigma^2$ and estimating σ^2 first, we estimate Q directly by calibrating the plug-in estimator. This new perspective establishes a general machinery taking reasonably good initial estimators of β and Σ as inputs. As shown in expression (6), the flexibility of the calibrated estimator has proven useful in efficiently pooling additional information on Σ whereas the estimation method that was introduced in Verzelen and Gassiat (2018) cannot be directly extended to integrating the unlabelled data in the semisupervised setting.

In numerical studies, we have demonstrated that the effect of including unlabelled data is of great practical significance, where, in the case of dense Σ, the RMSE of the new proposed

CHIVE estimator is 60–70% of the size of estimators (16) without using the unlabelled data. See Table 2 in Section 6 for details.

Additionally, Verzelen and Gassiat (2018) focused on the estimation problem instead of confidence interval construction and hypothesis testing problems. In terms of technical details on estimation optimality, the results in Verzelen and Gassiat (2018) allowed for a more general regime $k \geqslant \sqrt{p}$ than corollary 3 but did not handle the optimality in the semisupervised setting and did not allow the signal strength M to grow with n and p.

2.4. Case II: known Σ

The general semisupervised results also shed light on another interesting setting where the design covariance Σ is known. In the semisupervised setting, the unlabelled data are used for estimating Σ, so the case of known Σ is an extreme case of the semisupervised setting with N taken as ∞. The estimator (11) can be modified as $\hat{Q}(\hat{\beta}, \Sigma, Z^{(2)}) = \hat{\beta}^{\mathrm{T}}\Sigma\hat{\beta} + 2\hat{\beta}^{\mathrm{T}}(1/n_2)\Sigma_{i=n_1+1}^{n} X_{i\cdot}^{\mathrm{T}}(y_i - X_{i\cdot}\hat{\beta})$. Similarly, the estimator that was proposed in equation (6) is changed to $\hat{Q}(\hat{\beta}, \Sigma) = \hat{\beta}^{\mathrm{T}}\Sigma\hat{\beta} + 2\hat{\beta}^{\mathrm{T}}(1/n)\Sigma_{i=1}^{n} X_{i\cdot}(y_i - X_{i\cdot}^{\mathrm{T}}\hat{\beta})$.

Corollary 4. Suppose that assumption 1 holds and $k \leqslant cn/\log(p)$ for some constant $c > 0$.

(a) For any estimator $\hat{\beta}$ depending on the first half-sample $(y^{(1)}, X^{(1)})$ and satisfying condition 1, then, with probability larger than $1 - \gamma(n) - C\{p^{-c} + \exp(-ct^2)\}$,

$$|\hat{Q}(\hat{\beta}, \Sigma, Z^{(2)}) - Q| \lesssim (t+1)\frac{\sigma\|\Sigma^{1/2}\beta\|_2}{\sqrt{n}} + \frac{k\log(p)}{n}\sigma^2. \tag{17}$$

(b) For any estimator $\hat{\beta}$ satisfying condition 1, then, with probability larger than $1 - \gamma(n) - C\{p^{-c} + \exp(-ct^2)\}$,

$$|\hat{Q}(\hat{\beta}, \Sigma) - Q| \lesssim t\frac{\|\Sigma^{1/2}\beta\|_2}{\sqrt{n}} + \left(\frac{\|\Sigma^{1/2}\beta\|_2}{\sigma} + 1\right)\frac{k\log(p)}{n}\sigma^2. \tag{18}$$

Through comparing expression (17) with expression (12) and expression (18) with expression (7), the uncertainty of estimating the design covariance matrix leads to the additional term $\beta^{\mathrm{T}}\Sigma\beta/\sqrt{(N+n)}$. By applying theorem 2, it can be shown that the upper bound in expression (17) leads to the optimal convergence rate $M/\sqrt{n} + k\log(p)/n$. The term $M^2/\sqrt{(N+n)}$ disappears because of the known design covariance matrix Σ.

3. Semisupervised confidence intervals for $\beta^{\mathrm{T}}\Sigma\beta$

In this section, we quantify the uncertainty of the CHIVE estimator that was proposed in Section 2 and then construct confidence intervals for $\beta^{\mathrm{T}}\Sigma\beta$ in the semisupervised setting.

3.1. Confidence interval construction

The main next step of confidence interval construction for Q is to estimate consistently the standard error $\sqrt{\{4\sigma^2\beta^{\mathrm{T}}\Sigma\beta + \rho\mathbb{E}(\beta^{\mathrm{T}}X_{1\cdot}X_{1\cdot}^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\Sigma\beta)^2\}/\sqrt{n}}$ of the limiting distribution that was established in result (8). Specifically, we estimate $4\sigma^2\beta^{\mathrm{T}}\Sigma\beta$ by $\hat{\phi}_1$, ρ by $\hat{\rho} = n/(N+n)$ and $\mathbb{E}(\beta^{\mathrm{T}}X_{1\cdot}X_{1\cdot}^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\Sigma\beta)^2$ by $\hat{\phi}_2$, where $\hat{\phi}_1 = \hat{\sigma}^2\hat{\beta}^{\mathrm{T}}\hat{\Sigma}^{\mathrm{S}}\hat{\beta}$ and

$$\hat{\phi}_2 = \frac{1}{n+N}\sum_{i=1}^{n+N}(\hat{\beta}^{\mathrm{T}}X_{i\cdot}X_{i\cdot}^{\mathrm{T}}\hat{\beta} - \hat{\beta}^{\mathrm{T}}\hat{\Sigma}^{\mathrm{S}}\hat{\beta})^2,$$

with $\hat{\Sigma}^{\mathrm{S}}$ defined in equation (6). Then we propose the following confidence interval centred at $\hat{Q}$:

$$\mathrm{CI}(Z) = [(\hat{Q} - z_{\alpha/2}\hat{\phi})_+,\ \hat{Q} + z_{\alpha/2}\hat{\phi}], \qquad \hat{\phi} = \sqrt{\{(4\hat{\phi}_1 + \hat{\rho}\hat{\phi}_2)/n\}}, \tag{19}$$

where $z_{\alpha/2}$ is the upper $(\alpha/2)$-quantile of a standard normal distribution. The following theorem establishes the coverage and precision properties of $\mathrm{CI}(Z)$, where the length of the interval $\mathrm{CI}(Z) = (L(Z), U(Z))$ is defined as $\mathbf{L}\{\mathrm{CI}(Z)\} = U(Z) - L(Z)$.

Theorem 4. Suppose that assumptions 1 and 2 hold, $k \ll \min\{n/\{\log(N+n)\log(p)\}, \sqrt{n}/\log(p)\}$ and $\mathrm{SNR} \gg k\log(p)/\sqrt{n}$. For $\hat{\beta}$ and $\hat{\sigma}^2$ satisfying conditions 1 and 2 respectively, the confidence interval that is given in expression (19) satisfies

$$\lim_{n\to\infty} \mathbb{P}\{\beta^{\mathrm{T}}\Sigma\beta \in \mathrm{CI}(Z)\} \geqslant 1 - \alpha \tag{20}$$

$$\lim_{n\to\infty} \mathbb{P}[\mathbf{L}\{\mathrm{CI}(Z)\} \geqslant (1+\delta_0)\sqrt{\{4\sigma^2\beta^{\mathrm{T}}\Sigma\beta/n + \mathbb{E}(\beta^{\mathrm{T}}X_{1.}X_{1.}^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\Sigma\beta)^2/(N+n)\}}] = 0 \tag{21}$$

for any positive constant $\delta_0 > 0$.

The effect of the unlabelled data on the length of confidence interval is carefully characterized in equation (21), where the unlabelled data shrink part of the length of confidence interval, $\mathbb{E}(\beta^{\mathrm{T}}X_{1.}X_{1.}^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\Sigma\beta)^2/(N+n)$. This term corresponds to the uncertainty of estimating $\beta^{\mathrm{T}}\Sigma\beta$ in the oracle setting of known β. The most effective regime of integrating the unlabelled data is when the ratio

$$\frac{\mathbb{E}(\beta^{\mathrm{T}}X_{1.}X_{1.}^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\Sigma\beta)^2}{\sigma^2\beta^{\mathrm{T}}\Sigma\beta}$$

is not vanishing to 0. Otherwise, the dominating term in the length of expression (21) is $4\sigma^2\beta^{\mathrm{T}}\Sigma\beta/n$ and the additional unlabelled data are not helpful in this regime. In the numerical studies, we investigate how much shorter confidence intervals can be after integrating the unlabelled data. The lengths of confidence intervals in the semisupervised setting can be reduced to being as short as 30–40% of those in the supervised setting. See Table 2 in Section 6 for details.

The upper bound for confidence interval length established in expression (21) is further upper bounded by $\sigma\|\Sigma^{1/2}\beta\|_2/\sqrt{n} + \beta^{\mathrm{T}}\Sigma\beta/\sqrt{(N+n)}$, which matches the optimal convergence rate of estimation $M/\sqrt{n} + M^2/\sqrt{(N+n)}$ over the parameter space $\Theta(k, M)$ for $k \ll \sqrt{n}/\log(p)$ and $M \gg k\log(p)/\sqrt{n}$.

As shown in theorem 4, the validity of the proposed confidence interval (19) requires the condition that SNR is bounded away from zero by $k\log(p)/\sqrt{n}$. Although $k\log(p)/\sqrt{n}$ converges to 0 over the extreme sparse regime $k \ll \sqrt{n}/\log(p)$, it reveals the difficulty of constructing stable confidence intervals for $\beta^{\mathrm{T}}\Sigma\beta$ when SNR is at a local neighbourhood of 0. The next section will address the inference problem when SNR is at a local neighbourhood of 0.

3.2. Inference for weak signals

As discussed in Section 1, uncertainty quantification of $Q = \beta^{\mathrm{T}}\Sigma\beta$ is closely connected to other important statistical problems, including

(a) signal detection and global testing,
(b) prediction accuracy evaluation and
(c) confidence ball construction.

These applications provide a strong motivation for studying the inference problem for the explained variance under the settings of weak signals (i.e. $\mathrm{SNR} \lesssim k\log(p)/\sqrt{n}$). The main goal of this section is to discuss extensions of the proposed procedure to conduct statistical inference uniformly over different levels of signal strength, measured by SNR.

To begin with, we recall the reasoning for the non-uniformity assumption $\mathrm{SNR} \gg k\log(p)/\sqrt{n}$. This assumption is imposed such that the variance component of the CHIVE estimator dominates the bias component and in this case an asymptotic limiting distribution for the variance component is used to construct confidence intervals for the explained variance. Specifically, we discuss two possible solutions to remove this stringent assumption:

(a) to enlarge the confidence interval by an upper bound for the bias in Section 3.2.1;
(b) to increase the variance level by randomized calibration in Section 3.2.2.

3.2.1. *Bound the bias term*

One way to construct confidence intervals uniformly over all SNRs is to enlarge the estimated variance level that is defined in expression (19) to

$$\hat{\phi}^{\mathrm{E}} = \hat{\phi}^{\mathrm{E}}(y, X, \tau_0) = \sqrt{\left\{ \frac{1}{n} 4\hat{\sigma}^2(\hat{\beta}^{\mathrm{T}}\hat{\Sigma}^{\mathrm{S}}\hat{\beta} + \tau_0^2) + \frac{1}{(n+N)^2} \sum_{i=1}^{n+N} (\hat{\beta}^{\mathrm{T}} X_{i\cdot} X_{i\cdot}^{\mathrm{T}} \hat{\beta} - \hat{\beta}^{\mathrm{T}}\hat{\Sigma}^{\mathrm{S}}\hat{\beta})^2 \right\}}, \tag{22}$$

for some positive constant $\tau_0 > 0$. Then we construct the confidence interval as

$$\mathrm{CI}^{\mathrm{E}}(Z) = ((\hat{Q} - z_{\alpha/2}\hat{\phi}^{\mathrm{E}})_+, \hat{Q} + z_{\alpha/2}\hat{\phi}^{\mathrm{E}}), \tag{23}$$

where $z_{\alpha/2}$ is the upper $(\alpha/2)$-quantile of a standard normal distribution. The reason for adding the term $(1/n)4\hat{\sigma}^2\tau_0^2$ in the width (22) is that this additional term is an upper bound for the bias term in the regime $k \ll \sqrt{n}/\log(p)$. The following corollary establishes the coverage and the precision property of the enlarged confidence interval $\mathrm{CI}^{\mathrm{E}}(Z)$.

Corollary 5. Suppose that assumptions 1 and 2 hold, $k \ll \min\{n/\{\log(N+n)\log(p)\}, \sqrt{n}/\log(p)\}$ and $\tau_0 > 0$ is a positive constant. For $\hat{\beta}$ and $\hat{\sigma}^2$ satisfying conditions 1 and 2 respectively, then the confidence interval that is defined in equation (23) satisfies

$$\underline{\lim}_{n\to\infty} \mathbb{P}\{\beta^{\mathrm{T}}\Sigma\beta \in \mathrm{CI}^{\mathrm{E}}(Z)\} \geqslant 1-\alpha, \tag{24}$$

$$\lim_{n\to\infty} \mathbb{P}\left[\mathbf{L}\{\mathrm{CI}^{\mathrm{E}}(Z)\} \geqslant (1+\delta_0)\sqrt{\left\{ \frac{4\sigma^2(\beta^{\mathrm{T}}\Sigma\beta + \tau_0^2)}{n} + \frac{\mathbb{E}(\beta^{\mathrm{T}} X_{1\cdot} X_{1\cdot}^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\Sigma\beta)^2}{N+n} \right\}} \right] = 0 \tag{25}$$

for any positive constant $\delta_0 > 0$.

In contrast with the length of confidence interval in interval (21), the length in expression (25) is enlarged by the exact amount $4\sigma^2\tau_0^2/n$. In contrast with theorem 4, the inference is uniform over all levels of SNR at the expense of slightly longer confidence interval.

3.2.2. *Randomized calibration*

The construction in interval (23) still uses the CHIVE estimator as the centre and enlarges the constructed confidence interval. We introduce a randomized version of the CHIVE estimator as the new centre, where the main intuition is to increase the variance level through randomization such that the variance of this randomized estimator dominates its bias level. We generate random variables $u_i \sim^{\mathrm{IID}} N(0, \tau_0^2)$ for $1 \leqslant i \leqslant n$, independent of the observed data Z, and propose the following randomized calibrated estimator:

$$\hat{Q}^{\mathrm{R}} = \hat{Q}^{\mathrm{R}}(\hat{\beta}, \hat{\Sigma}^{\mathrm{S}}, u) = \hat{\beta}^{\mathrm{T}}\hat{\Sigma}^{\mathrm{S}}\hat{\beta} + 2\frac{1}{n}\sum_{i=1}^{n}(X_{i.}^{\mathrm{T}}\hat{\beta} + u_i)(y_i - X_{i.}^{\mathrm{T}}\hat{\beta}). \tag{26}$$

When there is no confusion, we use $\hat{Q}^{\mathrm{R}}$ to denote the estimator that is proposed in equation (26). In contrast with estimator (6), the calibration step in estimator (26) is involved with an additional term $2(1/n)\Sigma_{i=1}^{n} u_i(y_i - X_{i.}^{\mathrm{T}}\hat{\beta})$. If u_i is 0 instead of being generated as normal random variables in equation (26), the estimator $\hat{Q}^{\mathrm{R}}(\hat{\beta}, \hat{\Sigma}^{\mathrm{S}}, 0)$ is reduced to being exactly the same as $\hat{Q}(\hat{\beta}, \hat{\Sigma}^{\mathrm{S}})$ defined in expression (6). Since u_i in equation (26) are randomly generated normal random variables, this additional term approximately follows a normal distribution with mean 0 and variance $4\sigma^2\tau_0^2/n$. Even in the presence of weak signals, this additional term further enlarges the variance level of the calibrated estimator such that the bias level of the calibrated estimator is dominated by the corresponding variance level. The following theorem establishes the limiting distribution of the estimator $\hat{Q}^{\mathrm{R}}$ after randomized calibration.

Theorem 5. Suppose that assumption 1 holds, $k \ll \sqrt{n}/\log(p)$ and $\tau_0 > 0$ is a positive constant. For any estimator $\hat{\beta}$ satisfying condition 1, then

$$\sqrt{n}\frac{\hat{Q}^{\mathrm{R}} - Q}{\sqrt{\{4\sigma^2(\beta^{\mathrm{T}}\Sigma\beta + \tau_0^2) + \rho\mathbb{E}(\beta^{\mathrm{T}}X_{1.}X_{1.}^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\Sigma\beta)^2\}}} \xrightarrow{\mathrm{d}} N(0,1) \tag{27}$$

where $\rho = \lim n/(n+N)$.

In comparison with the limiting distribution (8) in theorem 1, theorem 5 requires no condition on SNR to establish the asymptotic limiting distribution while the variance level of the established normal distribution is enlarged by the amount $4\sigma^2\tau_0^2/n$. This additional variance term is a side effect of the randomized calibration. However, it enables a uniform inference procedure over all levels of SNR. Then we propose the confidence interval $\mathrm{CI}^{\mathrm{R}}(Z) = [(\hat{Q}^{\mathrm{R}} - z_{\alpha/2}\hat{\phi}^{\mathrm{E}})_+, \hat{Q}^{\mathrm{R}} + z_{\alpha/2}\hat{\phi}^{\mathrm{E}}]$, where $\hat{\phi}^{\mathrm{E}}$ is defined in expression (22). This confidence interval has the same length as that of interval (23) but different centres. The proposed estimator $\hat{Q}^{\mathrm{R}}$ enjoys the advantage of having an asymptotic normal distribution but it suffers from the same disadvantage as all randomized procedures, where the output is random even given the same data set. The following corollary characterizes the coverage and precision properties of $\mathrm{CI}^{\mathrm{R}}(Z)$.

Corollary 6. Under the same conditions as corollary 5, the coverage property in inequality (24) and precision property in equation (25) hold for the confidence interval $\mathrm{CI}^{\mathrm{R}}(Z)$.

Algorithm 1 in Table 1 summarizes the uncertainty quantification methods for $\beta^{\mathrm{T}}\Sigma\beta$.

We conclude this section with some additional comments. Compared with point estimation, construction of confidence intervals for the explained variance is a more challenging problem, mainly because we need to characterize the uncertainty of the estimator proposed. Specifically, accurate estimation of Q can be conducted uniformly over all levels of SNR whereas construction of confidence intervals uniformly over all levels of SNR requires much more effort. Another interesting observation is that inference for explained variance is different from that for linear functionals (Zhang and Zhang, 2014; van de Geer *et al.*, 2014; Javanmard and Montanari, 2014a, b; Cai and Guo, 2017), where the valid inference results for the latter do not depend on the magnitude of SNR.

4. Related semisupervised inference problem

The improvement due to integrating the unlabelled data is not just limited to the inference

Table 1. Algorithm 1: semisupervised uncertainty quantification for $\beta^{\mathrm{T}}\Sigma\beta$

Input: labelled data $\{y_i, X_{i\cdot}\}_{1\leqslant i\leqslant n}$ and unlabelled data $\{X_{i\cdot}\}_{n+1\leqslant i\leqslant n+N}$; $\tau_0 > 0$
Output: point estimator $\hat{Q} = \hat{Q}(y, X)$, $\hat{Q}^{\mathrm{R}} = \hat{Q}^{\mathrm{R}}(y, X, \tau_0)$ and variance estimator $\hat{\phi}^{\mathrm{E}} = \hat{\phi}^{\mathrm{E}}(y, X, \tau_0)$

Step 1: initialization—construct point estimator $\hat{\beta}$ and $\hat{\sigma}^2$ satisfying conditions 1 and 2; estimate Σ by $\hat{\Sigma}^{\mathrm{S}}$ defined in equation (6)
Step 2: calibration—estimate Q by the CHIVE estimator $\hat{Q}$ in equation (4) or its randomized version $\hat{Q}^{\mathrm{R}}$ in equation (26)
Step 3: uncertainty quantification—quantify the error of the proposed estimator by $\hat{\phi}^{\mathrm{E}}$ defined in equation (2)

problem for $\beta^{\mathrm{T}}\Sigma\beta$ but can also be obtained in the semisupervised inference for $\|\beta\|_2^2$. This unweighted quadratic functional is different from $\beta^{\mathrm{T}}\Sigma\beta$ as the covariance matrix Σ does not appear in the expression. Hence, it is even unclear whether the unlabelled data can be of any help. We introduce in this section a procedure integrating the unlabelled data and also carefully quantify the improvement by making use of the additional unlabelled data in the semisupervised setting.

The estimation of $\|\beta\|_2^2$ in the supervised setting was studied in Guo *et al.* (2019), where the error decomposition of the plug-in estimator $\|\hat{\beta}\|_2^2$ was established as $\|\hat{\beta}\|_2^2 - \|\beta\|_2^2 = 2\hat{\beta}^{\mathrm{T}}(\hat{\beta} - \beta) - (\hat{\beta} - \beta)^{\mathrm{T}}(\hat{\beta} - \beta)$. In Guo *et al.* (2019), the bias term $2\hat{\beta}^{\mathrm{T}}(\hat{\beta} - \beta)$ in the decomposition was estimated and hence the plug-in estimator $\|\hat{\beta}\|_2^2$ was corrected.

We illustrate here how the additional unlabelled data facilitate the bias correction step. We randomly split the labelled data (y, X) into two subsamples $(y^{(1)}, X^{(1)})$ with sample size n_1 and $(y^{(2)}, X^{(2)})$ with sample size n_2, where $n_1 \asymp n_2$. Let $\hat{\beta}$ denote an estimator of β that is produced by the first subsample $(y^{(1)}, X^{(1)})$ satisfying condition 1, where one example is the scaled lasso estimator (5) applied to $(y^{(1)}, X^{(1)})$. Then we construct a projection direction $\hat{u} \in \mathbb{R}^p$ and propose the estimator $\|\hat{\beta}\|_2^2$ as

$$\widehat{\|\beta\|_2^2} = \|\hat{\beta}\|_2^2 + 2\hat{u}^{\mathrm{T}} \frac{1}{n_2} \sum_{i=n_1+1}^{n} X_{i\cdot}(y_i - X_{i\cdot}^{\mathrm{T}}\hat{\beta}). \tag{28}$$

The unlabelled data are particularly useful in estimating the projection direction $\hat{u} \in \mathbb{R}^p$. The projection direction $\hat{u}$ is constructed as $\hat{u} = \hat{\Omega}\hat{\beta} = \Sigma_{l\in\mathrm{supp}(\hat{\beta})}\hat{\Omega}_{\cdot l}\hat{\beta}_l$ where $\hat{\Omega}_{\cdot l}$ is the constrained l_1-minimization for inverse matrix estimation estimator CLIME (Cai *et al.*, 2011) defined as

$$\hat{\Omega}_{\cdot l} = \arg\min \|m\|_1 \qquad \text{subject to } \|\tilde{\Sigma}m - e_l\|_\infty \leqslant \lambda_{\mathrm{S}} \tag{29}$$

with

$$\tilde{\Sigma} = \frac{1}{N+n_1}\left(\sum_{i=1}^{n_1} X_{i\cdot}X_{i\cdot}^{\mathrm{T}} + \sum_{i=n+1}^{n+N} X_{i\cdot}X_{i\cdot}^{\mathrm{T}}\right)$$

and $\lambda_{\mathrm{S}} \asymp \sqrt{\{\log(p)/(n_1+N)\}}$. The additional unlabelled data play a role in constructing the sample covariance matrix $\tilde{\Sigma}$ in estimator (29) and hence constructing the projection direction $\hat{u}$. The specific way of including the unlabelled data to improve the estimation accuracy of $\|\beta\|_2^2$ is different from that of $\beta^{\mathrm{T}}\Sigma\beta$, where the additional unlabelled data are used to estimate Σ directly in estimating $\beta^{\mathrm{T}}\Sigma\beta$ whereas the additional unlabelled data are used to estimate Σ^{-1} in estimating $\|\beta\|_2^2$. However, the high level idea is the same, i.e. making use of the flexibility of the calibrated estimator and properly incorporating the information about Σ that is contained in the unlabelled data.

Precision matrix estimation has been studied in the literature; see Cai *et al.* (2011) and the references therein. We restrict attention to $\hat{\Omega}$ satisfying the following condition.

Condition 3. The estimator $\hat{\Omega}$ satisfies $\mathbb{P}[\|\hat{\Omega}-\Omega\|_2 \gtrsim C_\Omega s\sqrt{\{\log(p)/(N+n)\}}] \geqslant 1-\gamma_1(N+n)$ where $\gamma_1(N+n)\to 0$, $s=\max_{1\leqslant l\leqslant p}\|\Omega_{l\cdot}\|_0$ and C_Ω is a constant depending on $\|\Omega\|_{L_1}$.

The CLIME estimator $\hat{\Omega}=(\hat{\Omega}_{\cdot 1}\ \hat{\Omega}_{\cdot 2}\ldots\hat{\Omega}_{\cdot p})$ with $\hat{\Omega}_{l\cdot}$ constructed in definition (29) is shown to satisfy condition 3 under certain regularity conditions. See the exact statement in Cai *et al.* (2011). We show in the following theorem that, with a sufficiently large amount of unlabelled data, the inference results for the semisupervised setting are distinguished from those in the supervised data.

Theorem 6. Suppose that assumption 1 holds, $k\leqslant cn/\log(p)$ for some constant $c>0$ and $c_0\leqslant \lambda_{\min}(\Omega)\leqslant\lambda_{\max}(\Omega)\leqslant C_0$ for some positive constants $C_0\geqslant c_0>0$. Suppose that $\hat{\beta}$ satisfies condition 1 and $\hat{\Omega}$ satisfies condition 3. Under the sample size condition $N+n\gg C_\Omega^2 k\{s\log(p)\}^2$, then, with probability larger than $1-\gamma(n)-C\{p^{-c}+\exp(-ct^2)\}-\gamma_1(N+n)$,

$$|\widehat{\|\beta\|_2^2}-\|\beta\|_2^2|\lesssim\sigma\frac{\|\beta\|_2}{\sqrt{n}}+k\frac{\log(p)}{n}\sigma^2. \tag{30}$$

In addition, if $(1/\sigma)\|\beta\|_2\gg k\log(p)/\sqrt{n}$ and ϵ_i are IID Gaussian random variables, then

$$\sqrt{\{n/(\sigma^2 V)\}}(\widehat{\|\beta\|_2^2}-\|\beta\|_2^2)\xrightarrow{\mathrm{d}}N(0,1),\qquad V=4\sum_{i=n_1+1}^{n}(\hat{u}^{\mathrm{T}}X_{i\cdot})^2/n_2^2. \tag{31}$$

The limiting distribution in expression (31) leads to the confidence interval construction

$$\mathrm{CI}_{\|\beta\|_2^2}=(\widehat{\|\beta\|_2^2}-z_{\alpha/2}\hat{\sigma}\sqrt{V},\widehat{\|\beta\|_2^2}+z_{\alpha/2}\hat{\sigma}\sqrt{V})$$

where $\widehat{\|\beta\|_2^2}$ is defined in equation (28), V is defined in expression (31) and $\hat{u}=\Sigma_{l\in\mathrm{supp}(\hat{\beta})}\hat{\Omega}_{\cdot l}\hat{\beta}_l$.

A few remarks are in order for the semisupervised inference for $\|\beta\|_2^2$. The results that were established in Guo *et al.* (2019) showed that the optimal rate for estimating $\|\beta\|_2^2$ in the supervised setting is

$$\frac{\sigma\|\beta\|_2}{\sqrt{n}}+(1+\|\beta\|_2)k\frac{\log(p)}{n}\sigma^2.$$

In contrast, the term

$$\|\beta\|_2 k\frac{\log(p)}{n}\sigma$$

disappears in the rate of convergence (30) by efficiently incorporating the unlabelled data. The improvement varies across different signal strengths, where the reduction in RMSE is limited if the signal strength $\|\beta\|_2$ is small but is significant if $\|\beta\|_2$ is large. Although integrating the unlabelled data is useful in reducing the RMSE for estimating both $\beta^{\mathrm{T}}\Sigma\beta$ and $\|\beta\|_2^2$, it is interesting to observe that the improvement by incorporating the unlabelled data is different, where, for estimating $\beta^{\mathrm{T}}\Sigma\beta$, part of the variance component is reduced but, for estimating $\|\beta\|_2^2$, the bias component is reduced by

$$\|\beta\|_2 k\frac{\log(p)}{n}\sigma.$$

More interestingly, when the size of the unlabelled data is sufficiently large and the spectrum of Σ is bounded away from 0 and ∞, the rate of estimating $\|\beta\|_2^2$ in expression (30) coincides with that of estimating $\beta^{\mathrm{T}}\Sigma\beta$ in expression (17).

Theorem 6 requires an additional sample size condition for the unlabelled data: $N+n \gg C_\Omega^2 k\{s\log(p)\}^2$. The general results for any $N \geqslant 0$ are given in section A in the on-line supplementary material.

The additional unlabelled data are not just useful in improving the estimation accuracy but are also useful in confidence interval construction. The specific effect is different from that for $\beta^{\mathrm{T}}\Sigma\beta$; the confidence interval for $\beta^{\mathrm{T}}\Sigma\beta$ is shortened as in interval (21) whereas the length of confidence interval $\mathrm{CI}_{\|\beta\|_2^2}$ is not shortened in terms of order of magnitude. However, the additional unlabelled data significantly weaken the model complexity and sample size condition for establishing the limiting distribution, where the sufficient condition for the supervised setting is $(1/\sigma)\|\beta\|_2 \gg k\log(p)/\sqrt{n}$ and $k \ll \sqrt{n}/\log(p)$. Corollary 6 has shown that the condition $k \ll \sqrt{n}/\log(p)$ is not needed if there is a sufficient amount of unlabelled data.

5. Statistical applications

In this section, we apply the inference procedure related to the CHIVE estimator to tackle several important statistical problems.

5.1. Application 1: signal detection and global testing

Signal detection is of great importance in statistics and related scientific applications and the detection problem in high dimensional linear regression has been studied in Arias-Castro *et al.* (2011) and Ingster *et al.* (2010). The inference procedure that is stated in algorithm 1 has profound implications on signal detection and the general global testing in high dimensional linear regression. We consider the global hypothesis testing problem $H_0 : (\beta-\beta^{\mathrm{null}})^{\mathrm{T}}\Sigma(\beta-\beta^{\mathrm{null}}) = 0$ *versus* $H_1 : (\beta-\beta^{\mathrm{null}})^{\mathrm{T}}\Sigma(\beta-\beta^{\mathrm{null}}) > 0$, which includes signal detection as a special case with $\beta^{\mathrm{null}} = 0$. We apply algorithm 1 with a given $\tau_0 > 0$ and obtain the point estimator $\hat{Q}^{\mathrm{R}}(y - X\beta^{\mathrm{null}}, X, \tau_0)$ and its standard error estimator $\hat{\phi}^{\mathrm{E}}(y - X\beta^{\mathrm{null}}, X, \tau_0)$. Then we propose the detection procedure, with type I error controlled at $\alpha \in (0,1)$ as $D(\tau_0) = \mathbf{1}\{\hat{Q}^{\mathrm{R}}(y - X\beta^{\mathrm{null}}, X, \tau_0) \geqslant \hat{\phi}^{\mathrm{E}}(y - X\beta^{\mathrm{null}}, X, \tau_0)z_\alpha\}$. Define the null parameter space $\mathcal{H}_0 = \{\theta = (\beta^{\mathrm{null}}, \Sigma, \sigma) : 1/M_1 \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant M_1, \sigma \leqslant M_2\}$ and the local alternative parameter space as

$$\mathcal{H}_1(\Delta) = \left\{\theta = (\beta, \Sigma, \sigma) : (\beta-\beta^{\mathrm{null}})^{\mathrm{T}}\Sigma(\beta-\beta^{\mathrm{null}}) = \frac{\Delta}{\sqrt{n}}, \right.$$
$$\left. \frac{1}{M_1} \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant M_1,\ \sigma \leqslant M_2 \right\}.$$

The following corollary establishes that $D(\tau_0)$ controls the type I error asymptotically and also establishes the asymptotic power function of the test proposed.

Corollary 7. Suppose that assumptions 1 and 2 hold, $\tau_0 > 0$ is a positive constant and the vector $\delta = \beta - \beta^{\mathrm{null}}$ satisfies the conditions that $\|\delta\|_0 \ll \min\{n/\{\log(N+n)\log(p)\}, \sqrt{n}/\log(p)\}$ and $\sqrt{\mathbb{E}(\delta^{\mathrm{T}}X_{1.}X_{1.}^{\mathrm{T}}\delta - \delta^{\mathrm{T}}\Sigma\delta)^2} \geqslant c_0\delta^{\mathrm{T}}\Sigma\delta$ for some positive constant c_0. Then, for any $\theta \in \mathcal{H}_0$, the type I error is controlled: $\overline{\lim}_{n\to\infty}\mathbb{P}_\theta\{D(\tau_0) = 1\} \leqslant \alpha$. For $\rho > 0$ and any $\theta \in \mathcal{H}_1(\Delta)$ with some positive constant $\Delta > 0$, then

$$\lim_{n\to\infty} \mathbb{P}_\theta\{D(\tau_0)=1\} = 1-\Phi^{-1}\left[z_\alpha - \frac{\Delta}{\sqrt{\{4\sigma^2(\delta^{\mathrm{T}}\Sigma\delta+\tau_0^2)+\rho\,\mathbb{E}(\delta^{\mathrm{T}}X_{1\cdot}X_{1\cdot}^{\mathrm{T}}\delta-\delta^{\mathrm{T}}\Sigma\delta)^2\}}}\right]. \quad (32)$$

The assumptions of corollary 7 are the same as those of corollary 6 from the perspective that the conditions that are imposed on β in corollary 6 are now imposed on the difference vector $\delta=\beta-\beta^{\mathrm{null}}$. One sufficient condition for the difference vector δ to be sparse is that both the true signal β and the null hypothesis β^{null} are sparse. Corollary 7 shows that, for any positive constant τ_0, $D(\tau_0)$ controls the type I error asymptotically. The asymptotic power of the test proposed is established in expression (32), where the additional unlabelled data prove useful in improving the power. See section D.2 of the on-line supporting information for an improvement in the numerical studies. For the finite sample performance, we have investigated how to choose the randomization level τ_0 in the simulation section. See section D.3 for the numerical performance.

5.2. Application 2: prediction accuracy assessment

Inference for explained variance has important applications to evaluating the out-of-sample prediction for a given sparse estimator $\check{\beta}$. To keep the notation consistent, we assume that $\check{\beta}$ is estimated on the basis of a training data set (X^0, y^0) and (X, y) are independent test data to evaluate its prediction accuracy. We start with computing the residual on the test data set $y-X\check{\beta}=X(\beta-\check{\beta})+\epsilon$. The out-of-sample prediction accuracy is defined as $\mathrm{PA}(\check{\beta})=\mathbb{E}_{x_{\mathrm{new}}}\{x_{\mathrm{new}}^{\mathrm{T}}(\check{\beta}-\beta)\}^2=(\check{\beta}-\beta)^{\mathrm{T}}\Sigma(\check{\beta}-\beta)$ and it is reduced to the explained variance for the residual model with outcome $r=y-X\check{\beta}$ and covariates X. Let $\hat{Q}^{\mathrm{R}}(r,X,\tau_0)$ and $\hat{\phi}^{\mathrm{E}}(r,X,\tau_0)$ denote the outputs of algorithm 1 with the labelled data $\{(r_i,X_{i\cdot})\}_{1\leqslant i\leqslant n}$ and unlabelled data $\{X_{i\cdot}\}_{n+1\leqslant i\leqslant n+N}$ as inputs. Then we propose the point estimator of $\mathrm{PA}(\check{\beta})$ as $\hat{Q}^{\mathrm{R}}(r,X,\tau_0)$ and the interval estimator for $\mathrm{PA}(\check{\beta})$ as

$$\mathrm{CI}_{\mathrm{PA}}(\check{\beta})=[(\hat{Q}^{\mathrm{R}}(r,X,\tau_0)-z_{\alpha/2}\hat{\phi}^{\mathrm{E}}(r,X,\tau_0))_+,\hat{Q}^{\mathrm{R}}(r,X,\tau_0)+z_{\alpha/2}\hat{\phi}^{\mathrm{E}}(r,X,\tau_0)]. \quad (33)$$

The following corollary establishes the convergence rate for the point estimator and the coverage and precision properties of the interval estimator.

Corollary 8. Suppose that assumptions 1 and 2 hold, $\tau_0>0$ is a positive constant and $c_0\leqslant\lambda_{\min}(\Omega)\leqslant\lambda_{\max}(\Omega)\leqslant C_0$, $\sigma\leqslant M_2$ for some positive constants $C_0\geqslant c_0>0$ and $M_2>0$. For any sparse estimator satisfying $\|\check{\beta}\|_0\leqslant C\|\beta\|_0$ and $C>0$ we have the following results.

(a) If $k\leqslant cn/\log(p)$ for some positive constant $c>0$, then, with probability larger than $1-\gamma(n)-C\{p^{-c}+\exp(-cN)+\exp(-ct^2)\}$,

$$|\hat{Q}^{\mathrm{R}}(r,X,\tau_0)-Q|\lesssim t\frac{\|\check{\beta}-\beta\|_2+\tau_0}{\sqrt{n}}+t\frac{\|\check{\beta}-\beta\|_2^2}{\sqrt{(N+n)}}+(\|\check{\beta}-\beta\|_2+1)\frac{k\log(p)}{n}. \quad (34)$$

(b) If $k\ll\min\{n/\{\log(N+n)\log(p)\},\sqrt{n}/\log(p)\}$ and $\sqrt{\mathbb{E}(\delta^{\mathrm{T}}X_{1\cdot}X_{1\cdot}^{\mathrm{T}}\delta-\delta^{\mathrm{T}}\Sigma\delta)^2}\geqslant c_0\delta^{\mathrm{T}}\Sigma\delta$ for $\delta=\beta-\check{\beta}$ and some positive constant c_0, then the confidence interval that is defined in expression (33) satisfies the coverage property $\underline{\lim}_{n\to\infty}\mathbb{P}\{\mathrm{PA}(\check{\beta})\in\mathrm{CI}_{\mathrm{PA}(\check{\beta})}\}\geqslant 1-\alpha$ and

$$\lim_{n\to\infty}\mathbb{P}\left[\mathbf{L}(\mathrm{CI}_{\mathrm{PA}(\check{\beta})})\geqslant C\left\{\frac{\|\check{\beta}-\beta\|_2+\tau_0}{\sqrt{n}}+\frac{\|\check{\beta}-\beta\|_2^2}{\sqrt{(N+n)}}\right\}\right]=0 \quad (35)$$

for some constant $C>0$.

Corollary 8 has shown that the precision of the confidence interval for the prediction accuracy is not just related to the sample sizes n and N, the sparsity k and the dimension p but is also

related to the accuracy of the evaluated estimator $\|\check{\beta}-\beta\|_2$. As characterized in expressions (34) and (35), the integration of the unlabelled data is useful in improving the estimation accuracy and confidence interval precision. See Section 6.1 and section D.4 in the on-line supporting information for the numerical performance.

5.3. Application 3: confidence ball construction

The prediction accuracy evaluation that was established in expression (33) can be used to construct a confidence ball for β. For the setting where $\lambda_{\min}(\Sigma)$ is known, then we have $\lambda_{\min}(\Sigma)\|\check{\beta}-\beta\|_2^2 \leqslant (\check{\beta}-\beta)^{\mathrm{T}}\Sigma(\check{\beta}-\beta)$ and construct the confidence ball for β as

$$\mathrm{CB}(\check{\beta}) = \left\{ \beta : \|\beta-\check{\beta}\|_2^2 \leqslant z_{\alpha/2}\frac{1}{\lambda_{\min}(\Sigma)}\hat{\phi}^{\mathrm{E}}(r, X, \tau_0) \right\} \tag{36}$$

As shown in expression (35), the radius of the confidence ball $\mathrm{CB}(\check{\beta})$ is upper bounded by $(\|\check{\beta}-\beta\|_2+\tau_0)/\sqrt{n}+\|\check{\beta}-\beta\|_2^2/\sqrt{(N+n)}$. To minimize the radius, we need to select the centre $\check{\beta}$ for the confidence ball (36) such that $\check{\beta}$ is sparse and $\|\check{\beta}-\beta\|_2$ is small. In the high dimensional literature, several penalized estimators have been shown to satisfy such properties, such as the lasso, scaled lasso and the Dantzig selector.

6. Simulation study

We carry out simulation studies in this section to demonstrate the numerical performance of the CHIVE estimator. Specifically, we illustrate the numerical improvement of pooling over the unlabelled data in Section 6.1; we compare the performance of the CHIVE estimator with the plug-in estimator in Section 6.2. Additional simulation results are postponed to section D in the on-line supplementary material.

We first introduce the general simulation set-up that is used for this section. We generate the high dimensional linear regression (1) with the dimension $p=800$ and the labelled data with sample size n and unlabelled data with sample size N. For the linear model (1), the covariates $\{X_{i\cdot}\}_{1\leqslant i\leqslant n}$ for the labelled data and also $\{X_{i\cdot}\}_{n+1\leqslant i\leqslant n+N}$ for the unlabelled data are generated in IID fashion to follow a multivariate normal distribution with mean 0 and covariance matrix $\Sigma\in\mathbb{R}^{800\times 800}$ and the errors $\{\epsilon_i\}_{1\leqslant i\leqslant n}$ are generated as an IID standard normal distribution.

6.1. Effect of pooling over additional unsupervised data

The focus of this section is to illustrate the improvement after integrating the unlabelled data in the semisupervised setting. We first consider the inference problem for $\beta^{\mathrm{T}}\Sigma\beta$ and then the out-of-sample prediction loss evaluation.

6.1.1. Inference for $\beta^{\mathrm{T}}\Sigma\beta$

We fix the labelled data sample size as $n=400$ and vary the unlabelled data sample size N across $\{2000, 6000, 20000\}$. We consider the following settings for the design covariance matrix Σ and high dimensional regression vector β.

(a) Across settings 1, 2 and 3, the regression coefficients are generated as $\beta_i = i/10$ for $1\leqslant i\leqslant 0$ and $\beta_i=0$ for $i\geqslant 11$; the covariance matrix Σ is generated as follows:
   (i) setting 1, $\Sigma_{ij}=0.5^{|i-j|}$;
   (ii) setting 2, $\Sigma_{ij}=0.35$ for $1\leqslant i\neq j\leqslant p$ and $\Sigma_{ii}=1$ for $1\leqslant i\leqslant p$;
   (iii) setting 3, $\Sigma_{ij}=0.7$ for $1\leqslant i\neq j\leqslant p$ and $\Sigma_{ii}=1$ for $1\leqslant i\leqslant p$.

(b) Across settings 4, 5 and 6, the regression coefficients are generated as $\beta_i = 1.5 \times 0.8^i$ for $1 \leqslant i \leqslant 800$. The covariance matrix Σ is generated as follows:
(i) setting 4, $\Sigma_{ij} = 0.5^{|i-j|}$;
(ii) setting 5, $\Sigma_{ij} = 0.35$ for $1 \leqslant i \neq j \leqslant p$ and $\Sigma_{ii} = 1$ for $1 \leqslant i \leqslant p$;
(iii) setting 6, $\Sigma_{ij} = 0.7$ for $1 \leqslant i \neq j \leqslant p$ and $\Sigma_{ii} = 1$ for $1 \leqslant i \leqslant p$.

Settings 1–3 correspond to the exact sparse case whereas settings 4–6 correspond to the approximate sparse case. Settings 1 and 4 correspond to the case of an approximated banded covariance matrix whereas settings 2, 3, 5 and 6 are about denser covariance matrices. The simulations are replicated over 1000 simulations. RMSE and the coverage and length of confidence intervals are presented in Table 2. Regarding RMSE, we observe that incorporation of unlabelled data reduces it significantly. The column under 'Ratio' reports the ratio of RMSE of the semisupervised method to that of the supervised method and RMSE of the semisupervised method is reduced to 33–57% of that of the supervised method, depending on the amount of the unlabelled data and also the structure on Σ. Since $X_{i\cdot}$ follows a multivariate Gaussian distribution, the variance component depending on the unlabelled data is expressed as $\mathbb{E}(\beta^{\mathrm{T}} X_{1\cdot} X_{1\cdot}^{\mathrm{T}} \beta - \beta^{\mathrm{T}} \Sigma \beta)^2/(N+n) = 2(\beta^{\mathrm{T}} \Sigma \beta)^2/(N+n)$. From setting 1 to setting 3, the value of $\beta^{\mathrm{T}} \Sigma \beta$ increases as Σ becomes denser and this explains why the effect of using the unlabelled data becomes more significant; the same phenomenon holds for settings 4–6.

In terms of constructed confidence intervals, both confidence intervals that were constructed in the semisupervised setting and the supervised setting have near 95% coverage whereas the confidence intervals that were constructed by using the unlabelled data have much shorter lengths. Specifically, we use Ratio to measure the ratio of the length of confidence interval in the semisupervised setting to that in the supervised setting and we observe that the length of confidence intervals can be reduced by as much as 70%.

The unlabelled data are not just useful in inference for $\beta^{\mathrm{T}} \Sigma \beta$, but also in prediction loss evaluation, which will be illustrated in what follows.

6.1.2. Prediction loss evaluation

We generate $\beta_i = i/5$ for $1 \leqslant i \leqslant 0$ and $\beta_i = 0$ for $i \geqslant 11$ and $\Sigma_{ij} = 0.5^{|i-j|}$. We fix the labelled data sample size as $n = 400$ and vary the unlabelled data sample size N across $\{2000, 6000, 20000\}$. We use these generated data (both labelled and unlabelled) to evaluate the out-of-sample prediction accuracy $(\hat{\beta}(\lambda) - \beta)^{\mathrm{T}} \Sigma (\hat{\beta}(\lambda) - \beta)$, where $\hat{\beta}(\lambda)$ is the lasso estimator based on independent training data $(X^{(0)}, y^{(0)})$ with sample size 300 with the tuning parameter λ:

$$\hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} \frac{\|y^{(0)} - X^{(0)} \beta\|_2^2}{2n_0} + \lambda \sum_{j=1}^{p} \frac{\|X_{\cdot j}^{(0)}\|_2}{\sqrt{n_0}} |\beta_j|.$$

Note that $(X^{(0)}, y^{(0)})$ is an independent copy of the labelled data (X, y). Specifically, we consider the three estimators $\hat{\beta}(\lambda_0)$, $\hat{\beta}(6\lambda_0)$ and $\hat{\beta}(10\lambda_0)$ with

$$\lambda_0 = \sqrt{\left(\frac{z_{1-1/(10p)}}{n_0} \right)}$$

and use the randomization level $\tau_0 = 2$ in terms of estimating this out-of-sample prediction accuracy.

The simulations are replicated over 1000 simulations and we report the numerical performance of both point and interval estimators of the corresponding prediction accuracy in Table 3. The observation is consistent with that for $\beta^{\mathrm{T}} \Sigma \beta$, where confidence intervals in both semisupervised

Table 2. Inference for $\beta^{\mathsf{T}}\Sigma\beta$ with $n = 400$ and $N = 2000, 6000, 20000$

Setting	*N*	*RMSE*			*Coverage*		*Length*		
		Semisupervised	*Supervised*	*Ratio (%)*	*Semisupervised*	*Supervised*	*Semisupervised*	*Supervised*	*Ratio (%)*
1	2000	0.420	0.733	57.2	0.950	0.942	1.598	2.769	57.7
	6000	0.370	0.751	49.2	0.950	0.949	1.388	2.791	49.7
	20000	0.341	0.732	46.6	0.933	0.940	1.291	2.777	46.5
2	2000	0.554	1.026	54.0	0.933	0.928	2.077	3.834	54.2
	6000	0.421	0.949	44.3	0.951	0.957	1.741	3.855	45.2
	20000	0.407	0.994	41.0	0.940	0.948	1.581	3.838	41.2
3	2000	0.813	1.612	50.4	0.950	0.958	3.213	6.539	49.1
	6000	0.642	1.654	38.8	0.960	0.946	2.510	6.530	38.4
	20000	0.559	1.597	35.0	0.942	0.956	2.148	6.509	33.0
4	2000	0.415	0.745	55.7	0.938	0.939	1.591	2.740	58.1
	6000	0.361	0.742	48.6	0.932	0.935	1.383	2.742	50.4
	20000	0.324	0.738	43.9	0.955	0.949	1.290	2.748	46.9
5	2000	0.589	1.088	54.2	0.953	0.968	2.329	4.462	52.2
	6000	0.496	1.149	43.2	0.934	0.939	1.909	4.447	42.9
	20000	0.465	1.181	39.4	0.936	0.935	1.713	4.441	38.6
6	2000	0.924	2.013	45.9	0.962	0.949	3.698	7.689	48.1
	6000	0.724	1.914	37.8	0.945	0.951	2.822	7.692	36.7
	20000	0.632	1.894	33.3	0.935	0.959	2.371	7.696	30.8

Table 3. Inference for the out-of-sample prediction accuracy $(\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta)$

Estimator	*Loss*	*N*	*RMSE*			*Coverage*		*Length*		
			Semisupervised	*Supervised*	*Ratio (%)*	*Semisupervised*	*Supervised*	*Semisupervised*	*Supervised*	*Ratio (%)*
$\hat{\beta}(\lambda_0)$	0.145	2000	0.269	0.279	96.3	0.910	0.898	0.896	0.898	99.8
		6000	0.270	0.281	96.3	0.918	0.902	0.895	0.897	99.8
		10000	0.262	0.273	96.0	0.924	0.910	0.895	0.897	99.8
$\hat{\beta}(6\lambda_0)$	1.818	2000	0.294	0.363	81.2	0.924	0.896	1.046	1.148	91.1
		6000	0.308	0.373	82.6	0.918	0.888	1.042	1.148	90.8
		10000	0.299	0.368	81.1	0.926	0.892	1.038	1.148	90.3
$\hat{\beta}(10\lambda_0)$	4.679	2000	0.365	0.548	66.5	0.930	0.928	1.318	1.841	71.6
		6000	0.378	0.553	68.3	0.934	0.902	1.291	1.839	70.2
		10000	0.362	0.551	65.8	0.920	0.916	1.267	1.841	68.8

and supervised settings have coverage but the semisupervised estimators are uniformly better than the supervised estimators in terms of both RMSE and the length of confidence interval. As observed in Table 3, across the three estimators $\hat{\beta}(\lambda_0)$, $\hat{\beta}(6\lambda_0)$ and $\hat{\beta}(10\lambda_0)$, the effect of unlabelled data is different. The effect of unlabelled data for estimating $\hat{\beta}(\lambda_0)$ is marginal whereas the effect of unlabelled data $\hat{\beta}(10\lambda_0)$ is much more significant, where RMSE and length of confidence interval can be reduced by 30%. This matches with the theory where, in the simulation setting of a Gaussian design, the unlabelled data reduce the term $\{(\hat{\beta}-\beta)^{\mathrm{T}}\Sigma(\hat{\beta}-\beta)\}^2/(N+n)$ and $(\hat{\beta}-\beta)^{\mathrm{T}}\Sigma(\hat{\beta}-\beta)$ is quite small (0.145) for $\hat{\beta}=\hat{\beta}(\lambda_0)$ and is much larger (4.679) for $\hat{\beta}=\hat{\beta}(10\lambda_0)$.

The semisupervised data are also useful for improving the power for signal detection. Since the detection power is improved by only around 5%, we defer the detailed results to the on-line supplementary material section D.2.

6.2. Comparison with other estimators

In what follows, we compare the CHIVE estimator with the plug-in estimator. We fix the size of unlabelled data at $N=2000$ and vary the labelled data sample size n across $\{200, 400, 600, 800, 1000\}$. The simulations are replicated over 500 simulations. We generate the design covariance matrix as $\Sigma_{ij}=0.5^{|i-j|}$ and the high dimensional regression vector β across the following three settings:

(a) *setting a*, β is generated with sparsity 10 where $\beta_j=j/10$ for $1\leqslant j\leqslant 10$ and $\beta_j=0$ for $j\geqslant 11$;
(b) *setting b*, β is generated with sparsity 50 where $\beta_j=j/50$ for $1\leqslant j\leqslant 50$ and $\beta_j=0$ for $j\geqslant 51$;
(c) *setting c*, β is generated as an approximate sparse vector with $\beta_j=0.5^{p-1}$.

We compare four different estimators, where 'CHIVE' and 'CHIVE.semi' denote the CHIVE estimator in the supervised setting and semisupervised setting respectively; 'Plugin' and 'Plugin.semi' denote the plug-in estimator $\hat{\beta}^{\mathrm{T}}\hat{\Sigma}\hat{\beta}$ in the supervised setting and semisupervised setting respectively. A numerical comparison is reported in Fig. 1. Across all three settings, it is observed that the proposed CHIVE estimator has achieved uniformly much better estimation accuracy than the plug-in estimators, in both supervised and semisupervised settings. This numerical observation demonstrates that the calibration step is useful in improving the estimation accuracy.

We also point out that the unlabelled data are useful only if they are incorporated in a proper way. Plugin.semi is another estimator also using the unlabelled data to estimate Σ, but it is only slightly better than the Plugin estimator. In contrast, together with the calibration machinery, CHIVE.semi uses the additional data in an efficient way and the corresponding RMSE is significantly reduced in comparison with the CHIVE estimator.

7. Real data application

In this section, we analyse a yeast data set that was reported in Bloom *et al.* (2013) and study how the genetic variants explain the colony sizes under various growth media. The goal is to estimate the heritability measures of colony sizes under various growth media, which represent the variance of the colony sizes explained by the genetic variants.

Bloom *et al.* (2013) investigated a large-scale genomewide association study of 46 quantitative traits based on 1008 *Saccharomyces cerevisiae* segregants crossbred from a laboratory strain and a wine strain. These quantitative traits are measures of end point colony size under 46 different

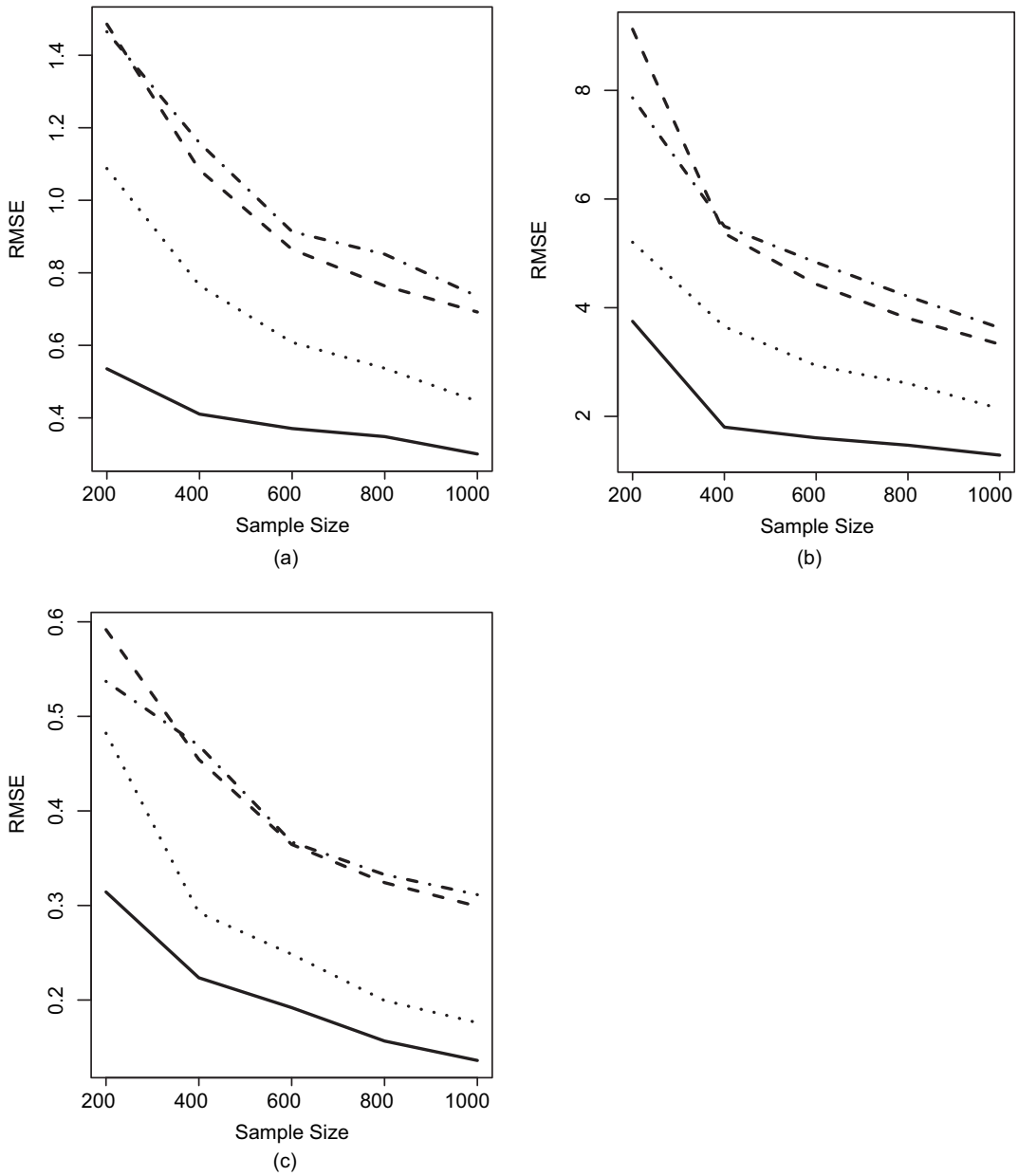

Fig. 1. RMSE of various estimators of $\beta^{\mathsf{T}}\Sigma\beta$ (the *x*-axis denotes the sample size and the *y*-axis denotes the RMSE of corresponding estimators; · · · · · · ·, RMSEs of the CHIVE estimator in the supervised setting; ———, RMSEs of the CHIVE estimator in the semisupervised setting; – – –, RMSE of the plug-in estimator in the supervised setting; · – · –, RMSE of the plug-in estimator in the semisupervised setting): (a) setting a (true value 9.42); (b) setting b (true value 49.47); (c) setting c (true value 2.9)

growth media, including hydrogen peroxide, cadmium chloride, calcium chloride, lactose, raffinose, sorbitol, yeast nitrogen base and yeast peptone dextrose. The genetic marker genotypes are coded as 1 or -1, according to which strain it comes from. A set of 11623 unique genotype markers of the 1008 segregants is measured. Since many of these markers are highly correlated and the corresponding codes are different in only several samples, Bloom *et al.* (2013) further selected a set of 4410 markers that are weakly dependent on the basis of the linkage disequilibrium information. All traits are normalized to have unit variance and hence the explained variance is a measure for heritability. Bloom *et al.* (2013) showed that the genetic variants are associated with many such trait values and highlighted the importance of addressing *missing heritability*. Bloom *et al.* (2013) pointed out one key reason for missing heritability as

> 'the undiscovered factors could have effects that are too small to be detected with current sample sizes, or even too small to ever be individually detected with statistical significance'.

We demonstrate that the CHIVE estimator has exactly addressed this concern of missing heritability. As reported in Table 4, we choose six traits out of the total 46 traits and observe that the CHIVE estimates are always larger than the corresponding plug-in estimates. This means that the calibration step adds back the missing heritability due to plugging in the lasso estimator, where the lasso estimator tends to ignore the genetic markers with small effects. The results for all 46 traits are reported in section E in the on-line supplementary material.

We also construct confidence intervals for heritability of all 46 traits and report part of the results in Table 4. Note that a proportion of the outcome variables for different growth media have missing values, with the proportion of missing ranging from 0.2% to 40.58%. This forms the semisupervised-type data naturally (note that the unlabelled data are of a smaller size than the labelled data in this specific example). After applying the proposed methods to analysing the corresponding outcomes, we have the following interesting observations:

(a) the heritability measures of the colony sizes under different growth media range from 0.3 to 0.8 and none of the confidence interval estimators contain zero; this means that the colony sizes under different growth media are strongly genetically heritable;
(b) the integration of the unlabelled data has shortened the length of the constructed confidence intervals; for example the length is shorter by around 3% for sorbitol (with 40.58% outcome missing), around 2% for raffinose (with 34.33% outcome missing) and around 1% for hydrogen peroxide (with 23.71% outcome missing).

8. Discussion

This paper studies statistical inference for the explained variance $\beta^{\mathrm{T}}\Sigma\beta$ in the semisupervised setting, which includes the supervised setting as a special case. By comparing the theoretical as well as the numerical results for the semisupervised and supervised settings, it is easy to see the significant contributions of the unlabelled data to the inference accuracy. In addition, the confidence interval constructed, using the idea of calibration, has been shown to be useful in tackling other important statistical applications, including signal detection and global testing, prediction accuracy evaluation and confidence ball construction. There remain a few open questions for future research.

Although the CHIVE estimator has been shown to achieve the optimal rates over the whole sparse regime $k \lesssim n/\log(p)$, construction of confidence intervals for $\beta^{\mathrm{T}}\Sigma\beta$ is only considered over the ultrasparse regime $k \ll \sqrt{n}/\log(p)$. Since neither the point nor the interval estimator requires prior knowledge of the exact sparsity level, they are referred to as adaptive estimation and the adaptive confidence interval respectively. However, it remains open whether it is possible

Table 4. Confidence intervals for heritability†

Medium	*Results for supervised*			*Results for semisupervised*			*Missing (%)*
	Plug-in	*CHIVE*	*CI*	*Plug-in*	*CHIVE*	*CI*	
Cadmium chloride	0.6240	0.7682 (0.0308)	[0.7077, 0.8286]	0.6215	0.7657 (0.0306)	[0.7058, 0.8256]	20.73
Calcium chloride	0.1807	0.3701 (0.0323)	[0.3068, 0.4333]	0.1785	0.3679 (0.0321)	[0.3050, 0.4308]	5.85
Hydrogen peroxide	0.2909	0.4835 (0.0380)	[0.4090, 0.5581]	0.2879	0.4806 (0.0375)	[0.4071, 0.5540]	23.71
Raffinose	0.3168	0.5105 (0.0410)	[0.4300, 0.5909]	0.3105	0.5041 (0.0399)	[0.4259, 0.5824]	34.33
Sorbitol	0.2968	0.4893 (0.0431)	[0.4049, 0.5737]	0.2864	0.4789 (0.0417)	[0.3972, 0.5606]	40.58
Yeast peptone dextrose	0.3754	0.5960 (0.0349)	[0.5275, 0.6645]	0.3761	0.5966 (0.0349)	[0.5282, 0.6651]	0.20

†The column indexed with 'Medium' represents the growth media for the yeast segregants. The three columns under 'supervised' correspond to the case of using only the labelled data, where the column indexed with 'Plug-in' represents the plug-in estimator, indexed with 'CHIVE' represents the CHIVE estimator and indexed with 'CI' represents the constructed confidence interval. Similarly, the three columns under 'Semisupervised' correspond to analysing the semisupervised-type data, i.e. also using the observations with missing outcome variables. The numbers in parentheses represent the standard errors of the CHIVE estimators proposed. The column indexed with 'Missing' represents the proportion of missing outcome for the corresponding media.

to construct adaptive confidence intervals over the moderate sparse regime $\sqrt{n}/\log(p) \lesssim k \lesssim n/\log(p)$. The possibility of adaptive confidence intervals for the general linear functional $\eta^{\mathrm{T}}\beta$ for $\eta \in \mathbb{R}^p$ has been studied in Cai and Guo (2017) and the technical tools that were developed in Cai and Guo (2017) can be useful to study the adaptive confidence intervals for $\beta^{\mathrm{T}}\Sigma\beta$.

Because of the emerging semisupervised data sets, it is of significant importance to propose procedures incorporating the unlabelled data efficiently and to study how the unlabelled data affect the statistical accuracy. This paper has studied both methodological and theoretical perspectives of the semisupervised statistical inference for the explained variance $\beta^{\mathrm{T}}\Sigma\beta$ and the unweighted quadratic functional $\|\beta\|_2^2$. However, it is largely unknown how these unlabelled data can facilitate the statistical inference problem for other quantities of interest, such as the general linear functional $\eta^{\mathrm{T}}\beta$ for some given $\eta \in \mathbb{R}^p$ and the variance level σ^2. These are interesting problems that are left for future research.

Acknowledgements

The research of Tony Cai was supported in part by National Science Foundation grant DMS-1712735 and National Institutes of Health grants R01-GM129781 and R01-GM123056. The research of Zijian Guo was supported in part by National Science Foundation grant DMS 1811857. The authors are grateful for the constructive and helpful comments from the Joint Editor, the Associate Editor and three referees.

References

Arias-Castro, E., Candès, E. J. and Plan, Y. (2011) Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Ann. Statist.*, **39**, 2533–2556.

Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98**, 791–806.

Bickel, P. J. and Ritov, Y. (1988) Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhya* A, **50**, 381–393.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Võ Lite, T.-L. and Kruglyak, L. (2013) Finding the sources of missing heritability in a yeast cross. *Nature*, **494**, 234–237.

Cai, T. T. and Guo, Z. (2017) Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity. *Ann. Statist.*, **45**, 615–646.

Cai, T. T. and Guo, Z. (2018a) Supplement to "Accuracy assessment for high-dimensional linear regression". *Ann. Statist.*

Cai, T. T. and Guo, Z. (2018b) Accuracy assessment for high-dimensional linear regression. *Ann. Statist.*, **46**, 1807–1836.

Cai, T. T. and Liu, W. (2011) Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Ass.*, **106**, 672–684.

Cai, T. T., Liu, W. and Luo, X. (2011) A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Ass.*, **106**, 594–607.

Cai, T. T. and Low, M. G. (2005) Nonquadratic estimators of a quadratic functional. *Ann. Statist.*, **33**, 2930–2956.

Cai, T. T. and Low, M. G. (2006) Optimal adaptive estimation of a quadratic functional. *Ann. Statist.*, **34**, 2298–2325.

Candès, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313–2351.

Chakrabortty, A. and Cai, T. (2018) Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.*, **46**, 1541–1572.

Collier, O., Comminges, L. and Tsybakov, A. B. (2017) Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.*, **45**, 923–958.

Donoho, D. L. and Nussbaum, M. (1990) Minimax quadratic estimation of a quadratic functional. *J. Complex.*, **6**, 290–323.

Efromovich, S. and Low, M. (1996) On optimal adaptive estimation of a quadratic functional. *Ann. Statist.*, **24**, 1106–1125.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42**, 1166–1202.
Gronsbell, J. L. and Cai, T. (2017) Semi-supervised approaches to efficient evaluation of model prediction performance. *J. R. Statist. Soc.* B, **80**, 579–594.
Guo, Z., Wang, W., Cai, T. T. and Li, H. (2019) Optimal estimation of genetic relatedness in high-dimensional linear models. *J. Am. Statist. Ass.*, **114**, 358–369.
Ingster, Y. I., Tsybakov, A. B. and Verzelen, N. (2010) Detection boundary in sparse regression. *Electron. J. Statist.*, **4**, 1476–1526.
van Iperen, E. P. A., Hovingh, G. K., Asselbergs, F. W. and Zwinderman, A. H. (2017) Extending the use of GWAS data by combining data from different genetic platforms. *PLOS One*, **12**, no. 2, article e0172082.
Janson, L., Barber, R. F. and Candès, E. (2017) Eigenprism: inference for high dimensional signal-to-noise ratios. *J. R. Statist. Soc.* B, **79**, 1037–1065.
Javanmard, A. and Lee, J. D. (2017) A flexible framework for hypothesis testing in high-dimensions. *Preprint arXiv:1704.07971*. Princeton University, Princeton.
Javanmard, A. and Montanari, A. (2014a) Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory. *IEEE Trans. Inform. Theory*, **60**, 6522–6554.
Javanmard, A. and Montanari, A. (2014b) Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, **15**, 2869–2909.
Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28**, 1302–1338.
Nickl, R. and van de Geer, S. (2013) Confidence sets in sparse regression. *Ann. Statist.*, **41**, 2852–2876.
Owen, A. (2012) Quasi-regression for heritability. Stanford University, Stanford. (Available from `http://statweb.stanford.edu/owen/reports/herit.pdf`.)
Raskutti, G., Wainwright, M. J. and Yu, B. (2010) Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.*, **11**, 2241–2259.
Sun, T. and Zhang, C.-H. (2012) Scaled sparse linear regression. *Biometrika*, **101**, 269–284.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Verzelen, N. and Gassiat, E. (2018) Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, **24**, no. 4B, 3683–3710.
Ye, F. and Zhang, C.-H. (2010) Rate minimaxity of the lasso and Dantzig selector for the lq loss in lr balls. *J. Mach. Learn. Res.*, **11**, 3519–3540.
Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.
Zhang, C.-H. and Zhang, S. S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B, **76**, 217–242.
Zhou, S. (2009) Restricted eigenvalue conditions on subgaussian random matrices. *Preprint arXiv:0912.4045*.
Zhu, J. and Bradic, J. (2017) A projection pursuit framework for testing general high-dimensional hypothesis. *Preprint arXiv:1705.01024*. University of Oregon, Eugene.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article.